Notes and Comments on S. Mallat's Lectures at Collège de France (2021)

Multiscale Models and Convolutional Neural Networks

J.E Campagne [*]

Janv. 2021; rév. 29 septembre 2023

[*] If you have any comments or suggestions, please send them to `jeaneric DOT campagne AT gmail DOT com`

Table des matières

1. Foreword

***Disclaimer**: What follows are my informal notes in French, translated into rough English, taken on the fly and reformatted with few personal comments ("NDJE" or dedicated sections). It is clear that errors may have crept in, and I apologize in advance for them. You can use the email address provided on the cover page to send me any corrections. I wish you a pleasant read.*

Please note that the Collège de France website has been redesigned. You can find all the course videos, seminars, as well as course notes not only for this year but also for previous years[1].

I would like to thank the entire Collège de France team for producing and editing the videos, without which the preparation of these notes would have been less convenient.

Also, note that S. Mallat[2] provides open access to chapters of his book *"A Wavelet Tour of Signal Processing"*, 3rd edition, as well as other materials on his ENS website.

This year, 2021, is the fourth in the cycle of S. Mallat's Data Science Chair: **Regularity, Approximation and Sparsimony**.

2. Lecture 13 Jan.

2.1 Introduction to the "Regularization, Approximation, Sparsity" Triangle

First, let's introduce this year's theme, which is "***Sparse Representations***." While in previous years, we studied *deep neural networks* with their applications, where we highlighted their empirical performance but also, from our perspective, a lack of mathematical support to truly understand them. This year, we return to a core aspect of **Data Processing**.

[1] https://www.college-de-france.fr/chaire/stephane-mallat-sciences-des-donnees-chaire-statutaire/events

[2] https://www.di.ens.fr/~mallat/CoursCollege.html

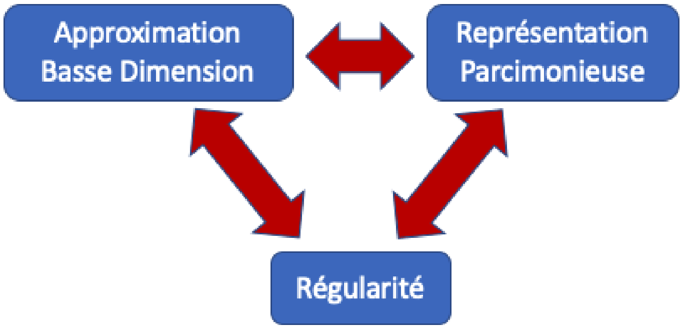

FIGURE 1 – The RAS triangle: "Regularity, Approximation, Sparsity."

If we denote $x(u) \in \mathbb{R}^d$ as the signal of interest (sound, image, time series, etc.), classical topics in **Signal Processing** include:

— **Approximation** of this signal. Indeed, we may want to transmit this signal with the fewest possible bits (Information Theory) to obtain $\tilde{x} \in \mathbb{R}^m$, and we try to quantify the *error* incurred (signal *distortion*), for example, through a norm $\|x-\tilde{x}\|$. Specifically, the approximation we are interested in is of **low dimension**, i.e., $m \ll d$, as we often want to **compress** the signal for transmission.

— **Denoising**. In this context, x is "contaminated" by noise/error ε, and we attempt to find a way to eliminate this nuisance. If the signal can be represented in a sparse form while the noise cannot, then we will see that we have a way to handle it and quantify the error.

— Finally, the **Inverse Problems** which we will revisit later.

In all of these themes, the goal is to recover x as cleanly as possible. Another significant field is **Analysis**, closely related to what is called **Statistical Learning**, which aims to answer the question: how to obtain y from x? In other words, we are **looking for a function** f, such that $y = f(x)$. Within this framework, we have themes like:

— **Classification**: e.g., determining if a given image is that of a cat, a boat, etc., or identifying a speaker as Mr. or Ms. X. In this case, y is a class indicator (integer).

— **Regression**, where in this case, y is a continuous quantity. For example, if x represents the distribution of atoms in a molecule, y is its minimum energy.

The function f is the underlying object, and we wonder if we can represent it with a minimal number of elements/parameters for efficient learning.

So, in these two main themes, we will consider the problem of *low-dimensional approximation*, which is related to *Sparse Representations*. In doing so, we will encounter a third concept: **Regularity**. These three concepts are intimately related (Fig. 1). For example, when considering sparse representations, the object of study (signal x or function f) is taken as a whole, and we want to represent it in a "basis" with very few non-zero coefficients. However, in practice, we cannot think of these representations without the notion of approximation. The choice to zero out coefficients is made with a criterion of approximation quality: minimizing the error incurred. And ultimately, when we discover sparse representations, we also discover forms of signal regularity and the underlying structure.

The interdependence of these three concepts is the subject of this year's course. Regarding applications, we will start with neural networks and then move on to signal processing. We will illustrate the "RAS" triangle in:

— The **linear domain**. Of course, we will encounter the entire harmonic analysis of Fourier. This is an essential foundation to master and is necessary to understand the subsequent material. We will address Sobolev regularities, and so on.

— And we will move on to the **non-linear domain** to understand why it is fundamental.

Keep in mind that every time we introduce new tools, we can revisit all the concepts in the RAS triangle: what are the structures highlighted, what are the approximation theorems, and the associated sparse representations.

It is clear that the theme of *sparsity* is not new. For instance, we can trace it back to *Occam's Razor*[3]. This philosophical principle also applies in science and, in essence, entails eliminating all explanations that are unnecessary. We can also trace it back to Aristotle, who considered one demonstration better than another if the former used fewer assumptions than the latter. We could continue to explore the use of this notion of sparsity in philosophy and science throughout the ages. This principle of minimal assumption is

[3] William of Ockham (c. 1285-1347): an English philosopher of the 14th century, a representative of nominalist scholasticism who criticized the possibility of a *demonstration* of divine existence. In this regard, he opposed the views of St. Thomas Aquinas, who synthesized Catholic theology and Aristotle's philosophy.

at the core of Newton's approach to building models, which progressively become more complex as our understanding of physical phenomena advances, rather than searching for Truth with a capital "T" [4]. What we can draw from this for the case at hand here and now is that we have "measurements" (x) that we need to explain using the most parsimonious representation systems possible.

A few more points to justify the use of sparsity. An empirical aspect, rather than a philosophical one, is that we avoid "overfitting": in essence, if the number of hypotheses is too large compared to the number of measurements, it becomes easier to provide an explanation. Another perspective concerns measurement errors: minimizing prediction error is a compromise between model error, *bias*, and statistical *variance*. In data compression, there is also a trade-off between signal quality and the number of bits of information used. Finally, sparsity can guide hypothesis selection to retain only those with the highest information density. This point will be addressed in the course through Information Theory and the concept of Entropy.

Finally, considering the prevalent aesthetic aspect, especially in mathematics, we might ask: can we make sparsity an absolute principle [5]? For instance, in biology, simplicity is not necessarily the rule. But in this context, it's also important to understand the situation in which the biological system evolves: does "simplicity" satisfy all the constraints the system faces (e.g., minimizing energy, adapting to potential predators, etc.)? So, we quickly realize that posing the question of simplicity/sparsity is only possible for isolated systems. In the course, we will only ask well-posed questions.

[4] Note: Isaac Newton was influenced both by Francis Bacon (1561-1626), who developed an empiricist theory of knowledge, and by Robert Boyle (1627-91), considered the father of modern natural philosophy. The "empirical philosophy" inspired by Bacon was in line with the thinking of the Royal Society of London. Therefore, while Newton's work is exceptional, it's not primarily due to the use of a revolutionary new method. However, explaining here the famous maxim "hypotheses non fingo" ("I do not feign hypotheses") would be too lengthy; this maxim is entwined with the theological aspects of his time.

[5] Note that the analytical philosophy emerging from the works of Gottlob Frege (1848-1925), Bertrand Russell (1872-1970), and Ludwig Wittgenstein (1889-1951) formulates science as a set of statements whose logical structure and meaning need to be found. In this context, sparsity plays a role in the selection of signs, for example.

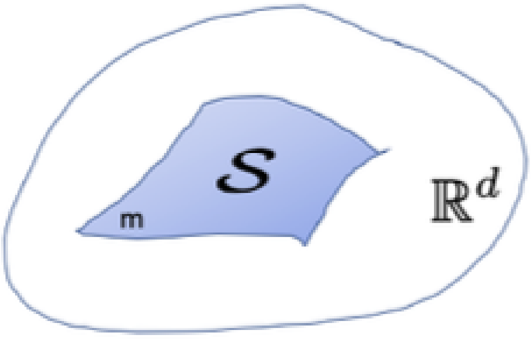

FIGURE 2 – The variety in which the data evolves, $\mathcal{S}$, is parameterizable by m coefficients.

2.2 Brief Illustrations of the RAS Triangle

2.2.1 Signal Processing

In the case of **compression**, we represent $x \in \mathbb{R}^d$ by $\Phi(x) \in \mathbb{R}^m$, aiming for $m \ll d$, meaning that the information in the message x can be reduced to m bits. Implicitly, this suggests that the signal (the data) doesn't evolve randomly in $\mathbb{R}^d$ but rather on a manifold $\mathcal{S}$ that might be included in $\mathbb{R}^m$, or at least parameterized by m coefficients (Fig. 2). In a way, $\Phi(x)$ is a local coordinate of $x \in \mathcal{S}$. In this context, discovering structures within the signal/data helps.

Once we understand that the structure constrains x to evolve on $\mathcal{S}$, *noising* is essentially taking the signal out of the surface $\mathcal{S}$. An idea for **denoising** is to reproject $x+\varepsilon$ onto $\mathcal{S}$ (Fig. 3). Of course, there is a denoising error ε', but it is much smaller than ε thanks to this projection. Complications arise when the underlying space is not necessarily linear, in which case non-linear projection methods are needed. What we observe is that the smaller the space in which x evolves (i.e., the dimensionality of $\mathcal{S}$), the more effective noise removal becomes: the sparser and/or lower-dimensional the representation, the greater the efficiency. An aspect that brings us into the RAS triangle is that the surface $\mathcal{S}$ can only be a model, an approximation of the geometric locus of all x. Thus, once again, we encounter two types of errors: one on the type of model because the signal does not exactly evolve on $\mathcal{S}$, and the other on the projection, which leaves residual noise.

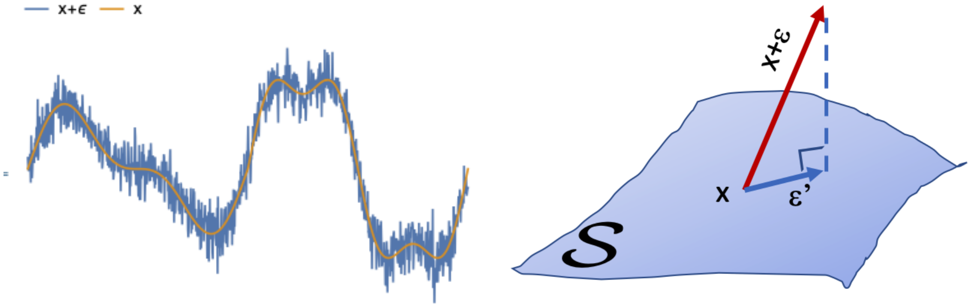

FIGURE 3 – Signal x and noise ε and a denoising process through orthogonal projection onto the surface $\mathcal{S}$.

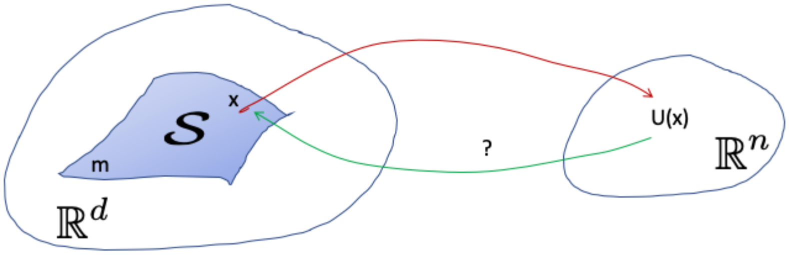

FIGURE 4 – Inverse Problems: Can we retrieve x from measurements on x, denoted $U(x)$?

The third type of problem, to some extent much more significant as mentioned, is the **Inverse Problems** (Fig. 4). Here, what we have is not x but a measurement of x, denoted $U(x) \in \mathbb{R}^n$, where n is the number of measured parameters, such that $n \leq d$:

$$x \in \mathbb{R}^d \xrightarrow{U} U(x) \in \mathbb{R}^n \qquad n \leq d \tag{1}$$

However, in this context, the operator U is not *invertible* (otherwise, the solution would be straightforward). To tackle this, we need prior information about x, particularly that x lies on a surface $\mathcal{S}$ contained in $\mathbb{R}^d$. Then, we can attempt inversion, but it requires that m, the number of parameters characterizing the manifold $\mathcal{S}$, be greater than n, the dimension of the space in which $U(x)$ operates. This is referred to as the inversion of the *restriction* of the operator U to $\mathcal{S}$. If the surface $\mathcal{S}$ is nonlinear, even if the operator U is linear (e.g., an average of measurements), the inversion is nonlinear. This necessitates the use of much more sophisticated algorithms and mathematics, even when dealing with a linear operator.

2.2.2 Statistical Learning

In this field, the question arises of finding a function f that provides the response y when given an input x: seeking f such that $y = f(x)$. This applies to problems of *classification* (where y is an integer or a vector of integers) or *regression* (where y is a real number or a vector of real numbers). Let's assume that $x \in [0,1]^d$ and $y \in \mathbb{R}$; the space in which the function f operates is colossal. We can make some assumptions, e.g., the conservation of energy, and then f belongs to the space of square-integrable functions:

$$L^2([0,1]^d) = \left\{ f / \int_{[0,1]^d} |f(x)|^2 dx < \infty \right\} \tag{2}$$

In this case, the space can be equipped with a quasi-Euclidean inner product, a pre-Hilbert space (infinite-dimensional[6]), allowing the definition of a norm between functions. However, the space $L^2([0,1]^d)$ is equally vast, and finding f requires techniques quite similar to Inverse Problems.

In particular, given a set of data $\{x_i\}_{i \leq n}$ and knowing the corresponding responses

[6] In finite dimensions, it's a Euclidean space

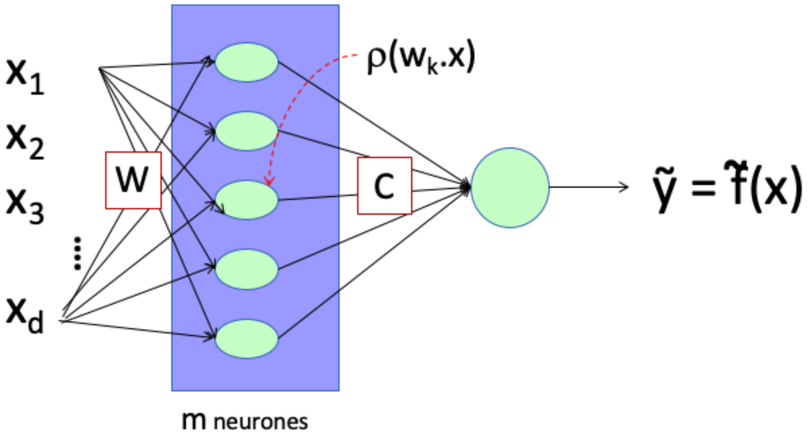

FIGURE 5 – One-hidden-layer neural network.

$\{y_i = f(x_i)\}_{i \leq n}$, we have a **Supervised Learning** scenario, also called an interpolation problem. However, strong assumptions about the class of functions for f must be made, and a sufficient number of samples (n) is needed to determine f in infinite dimensions. This is especially challenging when dealing with the *curse of dimensionality*[7].

From an algorithmic perspective, a one-hidden-layer neural network with m neurons (Fig. 5) has three distinct operations:

- A *linear operation* through the action of a matrix $W_{m,d}$, which can be seen as a scalar product over m vectors: $W_x = \{\langle x, e_p \rangle\}_{p \leq m}$
- A *pointwise non-linearity* ρ, such as a *rectifier* defined by $\rho(a) = \max(a, 0)$, and other choices are possible.
- Finally, a *linear classifier* C, which, in the case of regression, has dimensions of $(m, 1)$, or (m, K) for classification between K-classes.

Ultimately, we can write, introducing biases b_w and b_c at the level of the two linear combinations:

$$\boxed{\tilde{f}(x) = C\rho(W.x + b_w) + b_c} \tag{3}$$

[7] See the 2018 and 2019 courses, for example.

or in the case where y is a single number:

$$\tilde{y} = \tilde{f}(x) = \sum_p C_p \ \rho(\langle x, e_p \rangle + b_w) + b_c \tag{4}$$

If we ignore the biases for notation simplification, we realize that the response $\tilde{y}$ is a linear combination of elementary functions:

$$\tilde{y} = \sum_p C_p \ g_p(x) \qquad\qquad g_p(x) = \rho(\langle x, e_p \rangle) \tag{5}$$

In other words, to represent the function f, we have constructed *a simple linear model in relatively low dimension* (m) thanks to **elementary functions** $\{g_p\}_p$ based on the combination of a **scalar product** and a **non-linearity**.

For classification with K classes, y is a class label (e.g., $y = 1, \ldots, K$), and what we seek to approximate is $\log p(y|x)$ (the logarithm of the probability of y given x). This allows us to use the *Bayesian classifier*, which says that the best choice for y is the one for which the probability $p(y|x)$ is highest. One can view this type of problem as K regression problems, each followed by a max operation to obtain y. So, initially, we do not distinguish between a pure regression problem and a classification problem.

In conclusion, learning with a neural network traverses the RAS triangle: understanding how many neurons are needed according to the regularity of the function, what this will give in terms of the response's approximation, and perhaps discovering that if it works, it's because the matrices W and C are sparse, which is a form of sparsity. However, does one hidden layer suffice for the task at hand? In general, the answer is no, unless an enormous number of neurons in the hidden layer is required (see the *Universality Theorem of a One-Hidden-Layer Network* from the 2019 course). But in practice, there are cases where it works quite well. What does this mean? It means that f has **structure**! And why does f have structure? It somehow responds to the question of **regularity** in f.

When dealing with cases where one-hidden-layer networks do not work, we turn to **deep neural networks**, and this requires stepping out of the linear framework. Indeed, a deep network can be visualized as shown in Figure 6 and can be written as a **cascade of**

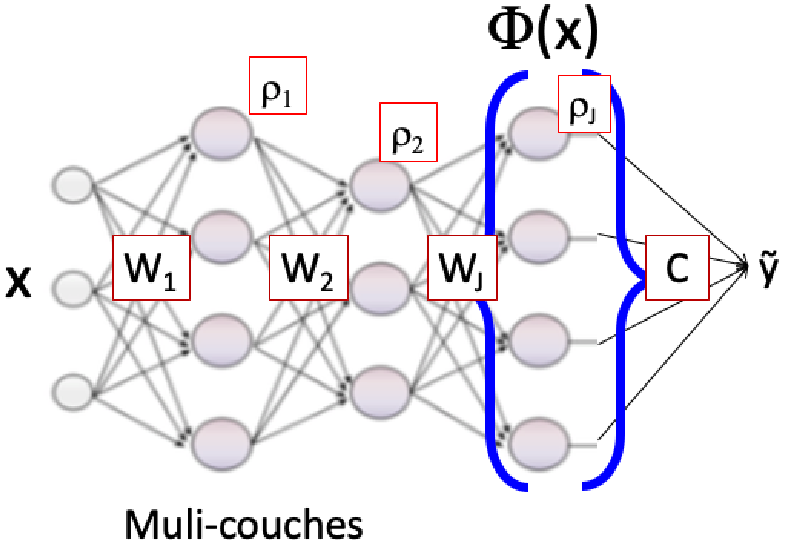

FIGURE 6 – Multi-layer network.

operators (ignoring biases here):

$$\boxed{f(x) = C\ \rho_J W_J \dots \rho_2 W_2\ \rho_1 W_1 x = C\Phi(x)} \tag{6}$$

whose result is **highly nonlinear**. Again, understanding deep neural networks involves going through the RAS triangle: what are the regularities of the learned functions? What structures are learned? Does sparsity play a role? Do we have theorems that guide judgment on the errors made? Etc. The major difficulty arises from asking these questions in a highly nonlinear and high-dimensional context. And we understand that to grasp the nonlinear, we must first understand what happens when we move from linear to nonlinear. Why is it necessary on one hand but a challenge worth taking on the other? One result is that in the nonlinear, we access sparser representations of much higher quality if they reflect the underlying regularity of the problem. There are cases where the nonlinear doesn't perform better than the linear, but generally, nonlinear works better, and different regularity classes are defined: the manifolds on which x evolves are curves.

2.3 Course Outline

The considerations in the previous section will be explored further:

— **The Linear**[8]: We will examine approximations of x (data/signal processing) or f (learning/analysis) by projections into linear spaces. Immediately, the first tool we will encounter is **Fourier Analysis** (Harmonic Analysis) as soon as we have some structure. In this framework, the RAP triangle is completely understood. Regularity is considered from the perspective of the decay of Fourier coefficients (especially **Sobolev**, etc.). We have approximation theorems, which naturally lead to sparse representations.

However, in certain linear cases, we do not know the basis of the representation, so we turn to **Principal Component Analysis** (PCA). We will not revisit the algorithms, but we will show the connection with Fourier Analysis and its limitations. As for one-hidden-layer neural networks, we will revisit the **Universal Approximation Theorem**, which is not mysterious, and we will also examine its limitations in high dimensions. The representation systems we will use will mostly be orthonormal **bases**.

— **The Nonlinear**[9]: We will also see how to perform approximations in bases, especially the concept of **thresholding** (adaptive). However, we need to find "good" bases, and, in particular, the ones from Fourier or PCA analyses are very poor (they are not designed for this purpose). Thus, we will revisit **Multi-resolution Analyses** with **orthonormal Wavelet bases**. We will see that fast Wavelet Transform algorithms (the counterpart of FFT for Wavelets) strangely resemble the structure of deep neural networks. From there, we will revisit the entire RAP triangle with the same concepts as in linear cases but in a nonlinear framework. Sparse representations will be different, low-dimensional approximations will be done with different algorithms, leading to different regularity classes described in more general spaces than Sobolev spaces, namely **Besov spaces**[10], where signals, instead of being uniformly regular, can have singularities and are more complex[11]. With these tools, we can address,

[8] Note: There is a lot of material on this theme in previous courses (2018-20).

[9] Note: See previous courses regarding, for example, Wavelets.

[10] Named after Oleg Vladimirovich Besov (1933-), a Russian mathematician.

[11] The measure of singularities is the index of the space, and the Dirac delta function is a member of certain Besov spaces

for example, images with contours, which have structures.

— **Information Theory**. We reach this point when we want to relate the concepts of the RAP triangle to models. We will see this in the context of Compression because what matters there is the number of bits, not the number of parameters: the difference? A bit is, as the name suggests, a binary number 0 or 1, whereas a parameter is generally a real number that theoretically requires an infinite number of bits to encode. The underlying challenge is **stability** because we need to find stable approximations for small bit transmission errors, for example. Of course, we will see that the number of bits needed to encode information is related to **Entropy**. This is the basis of Claude Elwood Shannon's (1916-2001) theory.

We notice that in very high dimensions, processes concentrate in very small spaces compared to the initial space; these are called *typical* sets (a well-established term), and their size is given by Entropy. We will see how we can achieve optimal compression codes with applications. In particular, we will examine **image compression** codes with two standards: JPEG, which mainly uses Fourier bases, and JPEG2000, which uses Wavelet bases. The second application will be **denoising**, which is not only a practical problem but also allows us to identify the space in which x operates.

We will explore both linear and nonlinear aspects of denoising with underlying models: the **Bayesian approach** and the **minimax approach**. Briefly, for data representation, there is a purely *deterministic* approach[12] that imposes a *prior* that can be summarized as: we know that x belongs to a set $\Theta \in \mathbb{R}^d$. In this context, we can hope to have the smallest possible global error over the set Θ: we want to minimize the maximum error that can occur if x moves through the entire space Θ. This introduces the concept of `minimax`:

$$\min \max_{x \in \Theta} \tag{7}$$

Underlying Bayesian models are *probabilistic*, which may seem paradoxical because probability implies uncertainty. In fact, it's the opposite because having a probabilistic model means having a lot of information to build the probability that x is in a certain part of the space Θ ($p(x)$). However, in practice, we almost never have access to $p(x)$. This is why we use `minimax` models to obtain rigorous results because the

[12] Note: See a discussion on Bayesian vs. determinism in the 2019 course.

idea is to consider the "worst-case" scenario.

Before concluding this section, let's consider a point that needs attention. So far, we've talked about space in terms of surfaces and manifolds. In very high dimensions, e.g., $\mathbb{R}^d$ with $d \gg 1$, even the surface itself is of very high dimension, and mathematically, we characterize it as a *random process*. In any case, we are not in the scenario of a 3-dimensional space projecting the signal onto a 2-dimensional surface, and it's clear that properties in nearly infinite (or truly infinite) dimensions are not the same as in low dimensions.

3. Lecture 20 Jan.

In this session, we will get a glimpse of how the Regularity, Approximation, and Sparsity (RAP) triangle unfolds differently when we are in a **linear** context compared to a **non-linear** one. We will work with two types of entities: either **data** in a broad sense, denoted as $x(u)$ and indexed by u (for example, time in 1D, pixel positions in a 2D image, etc.), or a **function** f that answers the question $y = f(x)$. So, depending on the domain, the object for which we seek an approximation that benefits from a good sparse representation according to its regularity will be either x or f, and in each case, it is essential to clarify which object is under study.

That said, in most cases throughout the course, we will primarily work with $x(u)$ viewed as a function of u, and when it comes to neural networks, we will revert to using the notation f.

3.1 A Simple Problem (Linear Context)

Consider the regular function $x(u)$, whose graph is shown in Figure 7. The problem at hand is to represent this function with *the fewest parameters* possible. One approach that comes to mind is to regularly sample this function, denoted as $\{x(nT)\}_{n \leq N}$, and perform regular interpolation $\tilde{x}(u)$ between these values. The error in approximating x by

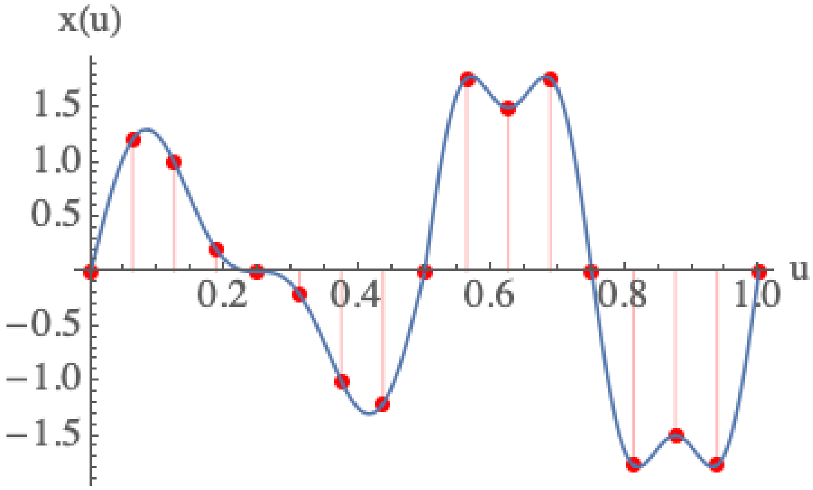

FIGURE 7 – A regular function $x(u)$ and regularly spaced sampling.

$\tilde{x}$ can be quantified using the quadratic difference (L2 norm):

$$\|x - \tilde{x}\|^2 = \int |x(u) - \tilde{x}(u)|^2 \, du \tag{8}$$

This type of approximation is *linear*. Please note that this does not mean using linear polynomials; it means that if we denote $\tilde{x}_1 \sim x_1$ to indicate that $\tilde{x}_1$ is an approximation of x_1, then by linear combination:

$$\left.\begin{array}{l} \tilde{x}_1 \sim x_1 \\ \\ \tilde{x}_2 \sim x_2 \end{array}\right\} \Rightarrow \lambda_1 \tilde{x}_1 + \lambda_2 \tilde{x}_2 \sim \lambda_1 x_1 + \lambda_2 x_2 \tag{9}$$

We are in a linear framework, and the approximation $\tilde{x}$ is characterized by $M = 1/T$ parameters (if the support is $[0, 1]$), so it resides in a space V_M of dimension $\dim(V_M) = M$. Thus, we have found a projection $\tilde{x}$ of the function x into this space V_M. However, we want to minimize the quadratic error Eq. 8:

$$\underset{\tilde{x} \in V_M}{\text{Min}} \ \|x - \tilde{x}\|^2 \tag{10}$$

Now, we know that the solution to this minimization problem leads to the orthogonal

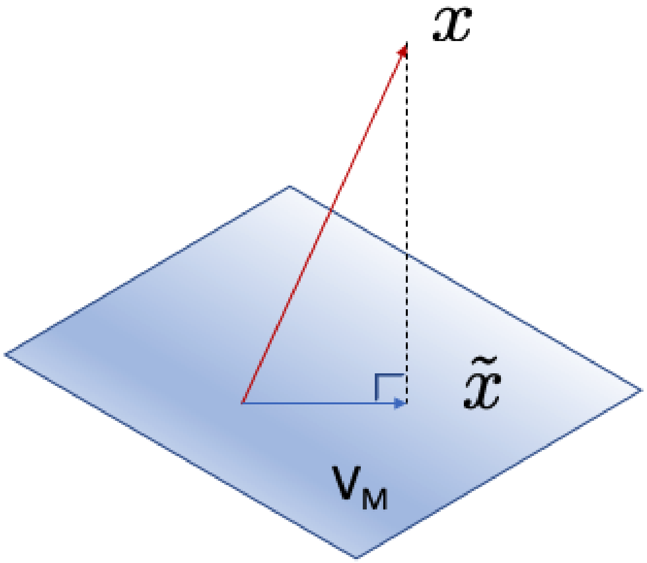

FIGURE 8 – Illustration of a linear approximation $\tilde{x}$ as the result of the orthogonal projection of x onto the linear space V_M.

projection of x onto the linear space V_M (Fig. 8):

$$\tilde{x} = P_{V_M}\, x \tag{11}$$

This is a general result in the linear case: **the approximation that minimizes the quadratic error is the orthogonal projection onto the considered linear space**. It's worth noting that, within the RAP triangle, we are currently on the side of Approximation.

How do we compute an orthogonal projection? One simple way is to use an orthonormal basis of the global space of dimension d, which always exists (recall that we are in finite dimension here). Let

$$\mathcal{B} = \{e_i\}_{i \leq d} \qquad \text{s.t.} \quad \langle e_i, e_j \rangle = \delta^K_{ij} = \delta^D[i-j] \tag{12}$$

We can rearrange the basis such that

$$\forall \tilde{x} \in V_M, \quad \tilde{x} = \sum_{i=1}^{M} \alpha_i e_i \tag{13}$$

The orthogonal projection of x onto V_M is then

$$\tilde{x} = \sum_{i=1}^{M} \langle x, e_i \rangle e_i \tag{14}$$

Now, how do we obtain bases that adapt to the fact that M can vary? First, we can work in the space of functions with support $[0, 1]$ and square-integrable $L^2([0, 1])$ functions. Note that we can also work in the space of functions with support in $\mathbb{R}$ and square-integrable $L^2(\mathbb{R})$ functions. These are spaces in which we can define an inner product between two functions:

$$\langle x, \tilde{x} \rangle = \int_{[0,1] \text{ or } \mathbb{R}} x(u)\tilde{x}^*(u)du \tag{15}$$

In these types of *infinite-dimensional* spaces, we construct an orthonormal basis $\mathcal{B} = \{e_n\}_{n\in\mathbb{N}}$ such that

$$\langle e_i, e_j \rangle = \delta[i - j] \tag{16}$$

and we have the following result on the error of projections:

$$\forall x \in L^2, \quad \lim_{M\to\infty} \left\| x - \sum_{i=1}^{M} \langle x, e_i \rangle e_i \right\|^2 = 0 \tag{17}$$

As in finite dimensions, the projection space of dimension M (finite) V_M is generated by the first M vectors of the basis $\{e_i\}_{i\leq M}$, and the error projects into the complementary space:

$$x - P_{V_M}x = \sum_{i>M} \langle x, e_i \rangle \; e_i \tag{18}$$

The advantage of using an orthonormal basis is that errors can be calculated easily. Indeed, it is this orthogonality that allows us to write:

$$\|x - P_{V_M}x\|^2 = (x - P_{V_M}x).(x - P_{V_M}x) = \sum_{i>M} |\langle x, e_i \rangle|^2 = \varepsilon_\ell \tag{19}$$

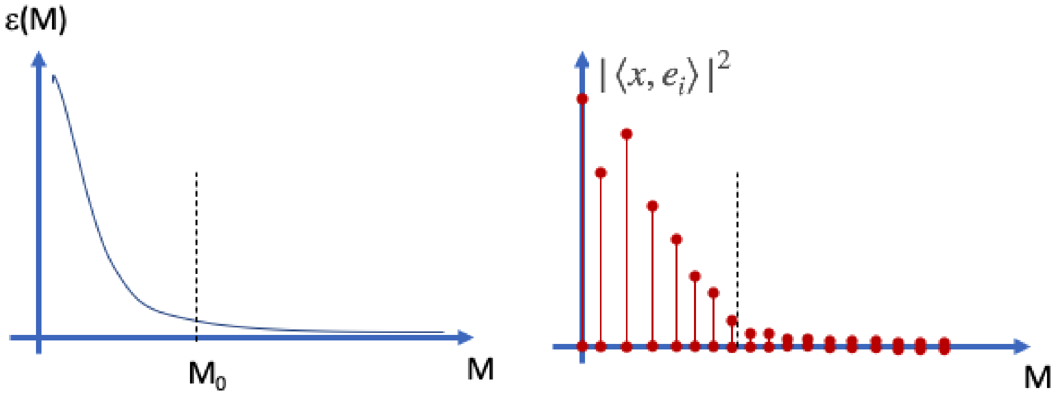

Figure 9 – (left): We would like the error of the (linear) approximation to decrease rapidly enough to set the threshold M_0 quite low. This is reflected in a constraint on the energy stored in the first coefficients (right).

We would like the approximation error, denoted as ε_ℓ (where ℓ stands for "linear"), to be as small as possible. Of course, we see from its previous expression that it depends on M (the larger M, the smaller it is). Moreover, we know that

$$\varepsilon_\ell(M) \xrightarrow[M\to\infty]{} 0 \tag{20}$$

Naturally, we would like (Fig. 9 left) the decrease to 0 to be as rapid as possible so that we can set a reasonably small threshold M_0, such that the remainder of the series $\sum_{i>M_0}$ is a small quantity. We can formulate the requirement as having the energy in all inner products beyond M_0 to be small. However, we know that for any orthonormal basis, there is the conservation of energy in x, which translates to:

$$\|x\|^2 = \sum_{i=1}^{\infty} |\langle x, e_i\rangle|^2 \tag{21}$$

So, it is necessary that **the energy of the first coefficients is the most significant** (Fig. 9 right), and thus, the decrease[13] of the sequence $(|\langle x, e_n\rangle|^2)_n$ to 0 must be rapid:

$$|\langle x, e_n\rangle|^2 \xrightarrow[n\to\infty]{\text{rapid}} 0 \tag{22}$$

[13] Note that the sequence is convergent because $\|x\|^2$ is finite.

What does this mean? First, we have highlighted the use of **a representation** in the orthonormal basis $\mathcal{B}$. The constraint on the inner products $|\langle x, e_n \rangle|$ to decrease rapidly implies **sparsity**, as in the representation, there are only a **small number of coefficients** that carry the signal's energy, if we want to make a **linear approximation**. **So performing an orthogonal projection is equivalent to setting coefficients beyond a rank M_0 to zero**. Now, the key to being able to make this approximation without too much error is that we have made an *a priori* assumption about the **regularity of the function** $x(u)$.

Here are some questions that can be asked in the **linear context**:

— What is this notion of regularity?

— What are the optimal approximation spaces V_M?

— And finally, what is the orthonormal basis $\mathcal{B}$ that achieves the best representation?

By answering these questions, we will discover Fourier analysis, Sobolev spaces, and encounter the problem of the curse of dimensionality.

3.2 A More Complex Problem (Non-Linear Context)

The example from the previous section, while generic, doesn't cover all scenarios, even for square-integrable functions. For instance, consider the 1-dimensional case shown in Figure 10. In 2D, in an image, you might encounter situations where from one pixel to its neighbor, you rapidly transition from one extreme to another on the grayscale due to the contours of an object against a uniform background. Often, **the "relevant" information lies in the discontinuities**.

Similar to the linear case, you can start with regular sampling at M points (left side of Fig. 10). You obtain a regular interpolation $\tilde{x}$ of the function x, and $\tilde{x}$ is the result of an orthogonal projection onto a linear space of dimension M (V_M). However, $\tilde{x}$ is chosen from among regular functions, so **errors are mainly concentrated where singularities occur**. Therefore, you need to change the locations of the samples, keeping M of them but concentrating them around the singularities. Thus, **sampling adapts** case by case according to x (right side of Fig. 10). The approximation that transitions from x to $\tilde{x}$ is then **non-linear**: if I have a function x_1 approximated by $\tilde{x}_1$, and another function x_2 approximated by $\tilde{x}_2$, the function $\alpha x_1 + \beta x_2$ is not approximated by $\alpha \tilde{x}_1 + \beta \tilde{x}_2$, because of the adaptation of sampling for each function (here M remains constant).

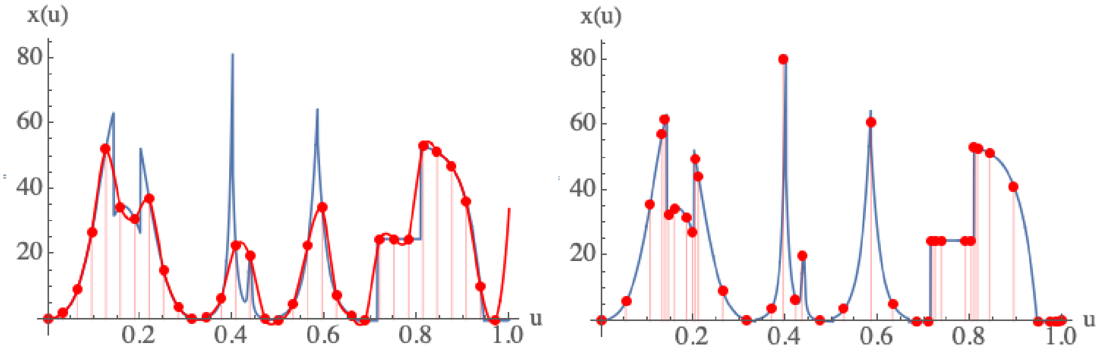

FIGURE 10 – Example of using regular sampling on the left and adaptive sampling on the right to better capture the singularities of the underlying function.

Alright, but how do you choose the right sampling? Certainly, we will try to locate the singularities, but how? And if we can do that, how do we distribute the M samples in proportion to the singularities? And we need to do this for any singular function (at least for a class of them). An approach that has been key to solving these problems is not very different from the one used in the previous problem. Let's start with an orthonormal basis $\mathcal{B}$; then we know that x can be decomposed as follows:

$$x = \sum_{i=0}^{\infty} \langle x, e_i \rangle e_i \tag{23}$$

Now, the M-parameter approximation $\tilde{x}$ can also be decomposed over the basis as follows:

$$\tilde{x} = \sum_{i \in \mathcal{S}} \langle x, e_i \rangle e_i \tag{24}$$

The point is that we will choose $\mathcal{S}$ with $|\mathcal{S}| = M$. **It's a partial sum, but not only with the first M coefficients.**

How will we choose them? What we want to do is minimize the quadratic error, and with the orthonormal basis, we can easily estimate this error:

$$\|x - \tilde{x}\|^2 = \sum_{i \notin \mathcal{S}} |\langle x, e_i \rangle|^2 \tag{25}$$

So the problem is to *minimize this error under the constraint* that we have M parameters.

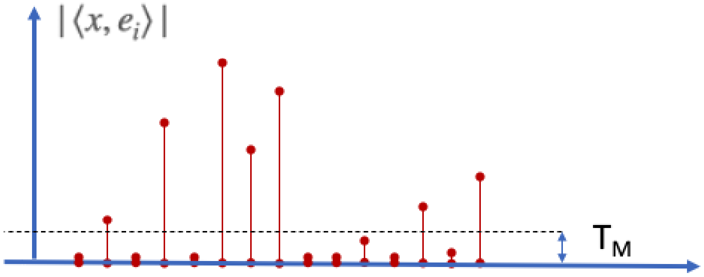

FIGURE 11 – Example of thresholding inner products to obtain adaptive sampling.

But for the error to be small, $\mathcal{S}$ must be the set of the M largest coefficients:

$$\mathcal{S}(x) = \{i \in \mathbb{N} \ / \ |\langle x, e_i \rangle| \text{ the top } M \text{ coefficients}\} \tag{26}$$

In fact, we should be able to order the inner products $|\langle x, e_i \rangle|$ from largest to smallest and take the first M. We'll see how to do this, but we can reformulate how to obtain $\mathcal{S}(x)$ by introducing a **threshold**, denoted as T_M, which ensures that we obtain M coefficients and for which (Fig. 11):

$$\mathcal{S}(x) = \{i \in \mathbb{N} \ / \ |\langle x, e_i \rangle| \geq T_M\} \tag{27}$$

The simplicity of implementing this scheme, despite it being non-linear, comes from the fact that the **basis $\mathcal{B}$ is orthonormal**. But which basis are we talking about? Because in the present case, it is about describing functions with singularities. By the way, we will see that the optimal basis in the linear context is the Fourier basis, but it is not at all capable of adapting to the singularities of the functions we are dealing with in non-linearity. So, we'll have to find something else, and we'll see that **Wavelet bases** fulfill the requirements. Now, is it worth it? Or in other words, **does the approximation error**, once we switch to adaptive sampling, decrease significantly? The answer to this question depends on **the regularity of the functions**. The notion of regularity, here in non-linearity, is much broader. And finally, functions like the one in Figure 10 are not as "irregular" as they may seem because despite their discontinuities here and there, there aren't many of them. Functions that are genuinely irregular, for example, are those that describe Brownian motion, where at every point, there is a singularity.

So, we will explore the RAP triangle in the non-linear context to discover new bases, new regularities, and new approximation theorems. Knowing that the transition from linear to non-linear occurs when moving from a one-layer network to a deep neural network.

In the following, we will first focus on the linear case because it is fundamental to understand non-linearity in the sense that we will recycle the same ideas, adapting them. And more precisely, we will approach the RAP triangle from the perspective of regularity.

3.3 What Is a Regular (Linear) Function?

From the perspective of **regularity**, we ask the following questions: can we construct approximations, build orthonormal bases, determine if they are optimal or not, and so on.

First, let's consider $x(u)$ as a time series. We can assess its regularity through its derivatives. If x is differentiable, we know it's already continuous, and its variations are smooth. Moreover, if its first derivative is bounded by a constant:

$$\left|\frac{dx(u)}{du}\right| \leq C \tag{28}$$

then we can say that the function $x(u)$ is rather regular. If we want even smoother functions, we need functions whose higher-order derivatives exist and are also bounded:

$$\forall k \leq n \quad \left|\frac{d^k x(u)}{du^k}\right| \leq C \tag{29}$$

This approach is the most natural. **Starting from this notion of regularity, we can unfold the entire RAP triangle**.

First, we ask the question: what is a derivative operator? We know that for $x \in L^2$, it yields the derivative dx/du, and this derivative is bounded. To understand this, we'll **diagonalize** it. But let's make a remark that relates to the 2020 course[14]: the derivative operator belongs to a broad class of **operators covariant/equivariant with respect to**

[14] NDJE: Section 6 of the 2020 course can be a supplement.

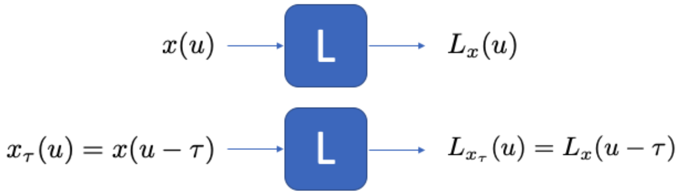

FIGURE 12 – Covariant/equivariant translation operator. The derivative operator is part of this type of operators.

translation:

$$\begin{aligned} g.x(u) = x(u-\tau) \Rightarrow D_u(g.x(u)) = \frac{dx(u-\tau)}{du} = \frac{dx}{du}(u-\tau) = g.D_u(x(u)) \\ \Rightarrow D_u(g.x) = g.D_u(x) \end{aligned} \tag{30}$$

(Note: an operator invariant under translation satisfies $f(g.x) = f(x)$). This covariant translation operator can be represented as shown in Figure 12. In the following, we will use the term *covariant* although it's actually *equivariance*. Note that this property is quite natural in signal processing. When transmitting a time series, we want operators that maintain the temporal sequence of values. So, a shift at the input should result in the same shift at the output. Regarding the derivative, this means that the derivative of a time-shifted signal is itself time-shifted (with the same time shift).

Now, the function $x(u)$ can be viewed as a sum of Dirac deltas:

$$x(u) = \int x(v)\delta(u-v)dv \tag{31}$$

Without going into distribution theory, we recall that the Dirac delta is the limit of functions whose "mass" concentrates at a point (integral equals 1 while having a support that tends to 0). If we apply an operator L covariant under translation, and if we have some level of regularity to interchange the integral and the action of operator L:

$$L.x(u) = \int x(v)L.[\delta(u-v)]dv = \int x(v)(L.\delta)(u-v)dv \tag{32}$$

In signal processing, the function $L.\delta(u) = h(u)$ is the **impulse response of the operator**

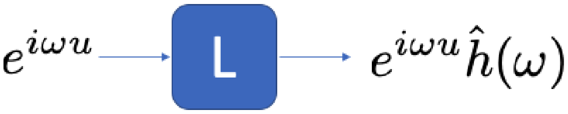

FIGURE 13 – The linear operator covariant under translation, denoted as L, has exponential functions $e^{i\omega u}$ as its eigenvectors, with associated eigenvalues equal to the values of its estimated Transfer Function at ω (cf. $\hat{h}(\omega)$).

L, and therefore, by covariance of L, we have:

$$\boxed{L.x(u) = \int x(v)h(u-v)dv = (x*h)(u) = (h*x)(u) = \int x(u-v)h(v)dv} \tag{33}$$

which translates to **the action of a linear operator covariant under translation L on x being a convolution of x with the impulse response of the operator**. So, the derivative operator belongs to the class of convolution operators. What are its eigenvectors? Let's take an oscillating exponential function $e^{i\omega u}$ at the input of the operator, and we get:

$$\boxed{L[e^{i\omega u}] = \int e^{i\omega(u-v)}h(v)dv = e^{i\omega u}\int h(v)e^{-i\omega v}dv = e^{i\omega u}\ \hat{h}(\omega)} \tag{34}$$

So, firstly, $e^{i\omega u}$ **is an eigenvector** of the linear operator L, and secondly, the associated eigenvalue is $\hat{h}(\omega)$, which is **the Fourier transform of the impulse response of the operator**, i.e., **the transfer function**. This can be illustrated as shown in Figure 13. Note that depending on the value of $\hat{h}(\omega)$, resonance phenomena can be highlighted. Therefore, an important point to remember is that **the Fourier Transform allows diagonalizing convolution operators**.

3.4 Fourier Analysis

Let's revisit some results on **Fourier Analysis** [15]. Fourier Analysis is a mathematical chapter that essentially concluded around the 1960s with the final convergence theorems for Fourier integrals and series.

[15] This fundamental topic is worth reviewing in previous year's courses as well.

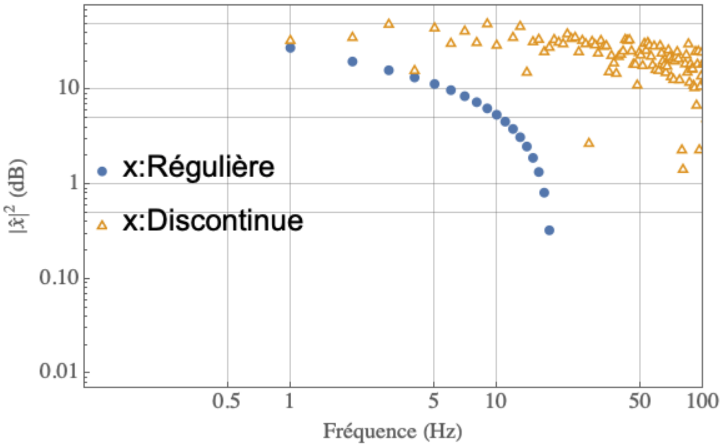

FIGURE 14 – Evolution of the Fourier coefficients of the two functions studied in the sections 3.1 (regular/linear framework) and 3.2 (discontinuous/nonlinear framework)

.

We define the Fourier Transform of x as follows:

$$\hat{x}(\omega) = \int_{-\infty}^{+\infty} x(u)e^{-i\omega u}du \tag{35}$$

To ensure the integral makes sense, we can restrict ourselves to functions in $L^1(\mathbb{R})$ for which:

$$\int |x(u)|du < \infty \tag{36}$$

Now, let's interpret $\hat{x}(\omega)$: it represents the result of the correlation between the function $x(u)$ and sinusoids whose ω sets the oscillation frequency. If the function x is regular, it oscillates slowly, so its coefficients $\hat{x}(\omega)$ are large for small ω. Conversely, if x has rapid variations, the "high-frequency" coefficients will be significant. Thus, the low frequencies indicate the regularity of a function. In other words, **the decay of Fourier coefficients reflects the regularity of the function**. An illustration is provided in Figure 14 for two functions studied in Sections 3.1 (regular/linear context) and 3.2 (discontinuous/nonlinear context). The regular function's Fourier coefficients converge to 0 much faster.

For the first theorem [16], we assume that the function $\hat{x}(\omega)$ is also in $L^1(\mathbb{R})$, meaning

[16] Proofs are provided in the course notes by S. Mallat.

it doesn't have too many high-frequency components.

> **Theorem 1** *If $\hat{x} \in L^1(\mathbb{R})$, then there is an inversion formula to recover the function x from its Fourier Transform:*
>
> $$x(u) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{x}(\omega) e^{i\omega u} d\omega \tag{37}$$

Note that the operation of recombining sinusoids by weighting them with Fourier coefficients is not straightforward. Because the sinusoids are spread out, and a very precise weighting is required to reconstruct, for example, a function that is mostly regular except for a small portion of its support (see, for example, Figure 22 of the 2020 course).

A consequence of this is:

$$|x(u)| = \frac{1}{2\pi} \left| \int_{-\infty}^{+\infty} \hat{x}(\omega) e^{i\omega u} d\omega \right| \leq \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{x}(\omega)| d\omega \tag{38}$$

Since $\hat{x} \in L^1(\mathbb{R})$, the right-hand side is bounded, and therefore, so is x. It has **bounded variations**, and

$$\|x\|_\infty = \sup_{u \in \mathbb{R}} x(u) \leq \frac{1}{2\pi} \int |\hat{x}(\omega)| d\omega \tag{39}$$

The second theorem follows from the intention to use the Fourier Transform for convolution:

> **Theorem 2**
>
> $$g(u) = (x * h)(u) \rightarrow \hat{g}(\omega) = \hat{x}(\omega)\hat{h}(\omega)$$

This means that the convolution product is diagonalized in Fourier space. Indeed, the convolution operator $Lx(u) = (x * h)(u)$ is linear and **covariant under translation** because

$$\begin{aligned} L[g_\tau.x(u)] = L[x(u-\tau)] = L[f(u)] &= \int f(v)h(u-v)dv \\ &= \int x(v-\tau)h(u-\tau-v+\tau)dv = \int x(v)h(u-\tau-v)dv = Lx(u-\tau) \\ &= g_\tau.(L_x(u)) \end{aligned}$$

so it is diagonalizable by complex exponentials as we showed earlier. Therefore,

$$g(u) = (x * h)(u) = Lx(u) = \frac{1}{2\pi} \int \hat{x}(\omega) L[e^{i\omega u}] d\omega = \int \hat{x}(\omega) \hat{h}(\omega) e^{i\omega u} d\omega$$

resulting in the identification of $\hat{x}(\omega)\hat{h}(\omega)$ as the Fourier coefficient $g(\omega)$.

The third (fundamental) theorem is the Plancherel formula, which reflects that the Fourier Transformation conserves angles (up to a constant factor):

> **Theorem 3**
>
> $$\langle x_1, x_2 \rangle = \int x_1(u) x_2^*(u) du = \frac{1}{2\pi} \int \widehat{x_1}(\omega) \widehat{x_2}(\omega) d\omega = \frac{1}{2\pi} \langle \widehat{x_1}, \widehat{x_2} \rangle \tag{40}$$
>
> *A consequence when $x_1 = x_2$ yields a relation (Parseval) which is an energy conservation:*
>
> $$\|x\|^2 = \|\hat{x}\|^2 \Leftrightarrow \int |x(u)|^2 du = \frac{1}{2\pi} \int |\hat{x}(\omega)|^2 d\omega \tag{41}$$

So, if $x \in L^2(\mathbb{R})$, then its Fourier Transform is also of finite energy, $\hat{x} \in L^2(\mathbb{R})$. **The FT preserves the space** $L^2(\mathbb{R})$, and we will mainly define it in this space.

Other properties of the Fourier Transform are useful for the following, and they are listed in Table 1.

3.5 The Derivative Operator: Sobolev Regularity

Let's return to the derivative operator, and first, let's discuss its eigenvectors. Since it is translation-covariant, as we studied in the previous section, its eigenvectors are $e^{i\omega u}$. The corresponding eigenvalue is simply $i\omega$, or in terms of the transfer function of the derivative operator:

$$\hat{h}_{d/du}(\omega) = i\omega \tag{42}$$

Now, if we want to say that the kth derivative of x is bounded, meaning:

$$\left\| \frac{d^k x(u)}{du^k} \right\|_\infty \leq C \tag{43}$$

FT	$\hat{f}(\omega) = \frac{1}{\sqrt{m}} \int f(x) e^{iq\omega x} dx$
Inverse FT	$f(x) = \frac{\sqrt{m}}{2\pi} \int \hat{f}(\omega) e^{-iq\omega x} dx$
Translation	$\widehat{f(x-h)}(\omega) = e^{iqwh} \hat{f}(\omega)$
Dilation ($s > 0$)	$\widehat{\frac{1}{s} f(\frac{x}{s})}(\omega) = \hat{f}(s\omega)$
Parity	$\widehat{f(-x)}(\omega) = \hat{f}(-\omega) = \widehat{f^*}^*(\omega)$
Average	$\hat{f}(0) = \frac{1}{\sqrt{m}} \int f(x) \mathrm{d}x$
Plancherel	$\int f(x) g^*(x) \mathrm{d}x = \frac{m}{2\pi} \int \hat{f}(\omega) \hat{g}^*(\omega) \mathrm{d}\omega$
Parseval	$\int \lvert f(x) \rvert^2 \mathrm{d}x = \frac{m}{2\pi} \int \lvert \hat{f}(\omega) \rvert^2 \mathrm{d}\omega$
Derivative	$\widehat{f^{(p)}(x)}(\omega) = (-iq\omega)^p \hat{f}(\omega)$
Convolution	$\widehat{f * g} = \sqrt{m} \hat{f}(\omega) \hat{g}(\omega)$
Multiplication by a Phase	$\widehat{e^{i\xi x} f(x)}(\omega) = \hat{f}(\xi/q + \omega)$

Table 1 – Examples of properties of the Fourier Transform expressed in a generic 1D formulation. S. Mallat uses ($m = 1, q = -1$), which also corresponds to the convention ($a = 1, b = -1$) in Mathematica; but in the literature, you may find formulations with $m = 1, 2\pi$ and $q = \pm 1, \pm 2\pi$.

then one way to proceed is to impose that the Fourier Transform of the kth derivative is integrable (see Eq. 39). However, knowing that:

$$\frac{d^k x(u)}{du^k} \xrightarrow{FT} (i\omega)^k \hat{x}(\omega) \tag{44}$$

this means that we must impose the condition:

$$\int |\omega|^k |\hat{x}(\omega)| d\omega = C \tag{45}$$

In other words, the previous condition in the Fourier domain *implies* the condition on the kth derivative in the real space. This reveals **a condition of regularity** for the function x, because if the Fourier coefficients decrease rapidly enough so that, when multiplied by ω^k, the integral remains finite, this constrains the variations of the kth derivative.

Can we establish an equivalence? To do this, let's use **Parseval/Plancherel**. We will replace the condition "the kth derivative is bounded" with the condition "the kth derivative is square-integrable", i.e.:

$$\int \left| \frac{d^k x(u)}{du^k} \right|^2 du = \frac{1}{2\pi} \int ||\omega|^k \hat{x}(\omega)|^2 d\omega = \frac{1}{2\pi} \int |\omega|^{2k} |\hat{x}(\omega)|^2 d\omega \leq \infty \tag{46}$$

So, there is an equivalence here. If the right-hand integral converges, then the kth derivative is square-integrable. This is **Sobolev regularity or differentiability**. In fact, this exponent can be extended to the case of a positive real number, and the condition then becomes:

$$s \in \mathbb{R}^+, \int |\omega|^{2s} |\hat{x}(\omega)|^2 d\omega \leq \infty \tag{47}$$

If the integral converges, we say that the function is *"s times"* differentiable with s a positive real number. This generalizes the notion of a derivative, and especially **in practical signal processing, we never actually compute the derivatives**, but rather, we go to the Fourier domain to study **the decay of coefficients**. Now, if the Fourier Transform decreases rapidly enough, we can set a lower-frequency cutoff and keep only a small number of coefficients, achieving **sparse representation**. However, one point to clarify is that, for now, if we remain in the continuous domain, applying a threshold to ω still leaves us with an infinite number of frequencies. The same goes for the calculation of Fourier integrals; we would need an infinite number of frequencies. However, the **goal of signal processing** is to **manipulate as few parameters as possible**.

3.6 Transition from Continuous to Discrete

One remark we can make is that in practice, the support of the signal $x(u)$ is finite, and we normalize it to be in $u \in [0, 1]$. In the background, if we need to extend it beyond, we consider $x(u)$ to be *periodic* (or extend it to a periodic function). Thus, we consider signals in $L^2([0, 1])$. We can redo all the previous Fourier analysis. The building blocks are the sinusoids, but this time with the constraint of having a period of 1, so $\omega = 2\pi n$ for all integers. We then have the following theorem:

> **Theorem 4**
>
> $$\mathcal{B} = \left\{ e_n(u) = e^{i2\pi nu}, \forall n \in \mathbb{Z} \right\} \tag{48}$$
>
> *$\mathcal{B}$ is an* ***orthonormal basis*** *for $L^2([0, 1])$, which is the (famous) result about Fourier series:*
>
> $$\forall x \in L^2([0,1]),\ x = \sum_{n\in\mathbb{Z}} \langle x, e_n \rangle e_n$$
>
> *with*
>
> $$\langle x, e_n \rangle = \int_0^1 x(u) e^{-i2\pi nu} du = \hat{x}(2\pi n)$$
>
> *which is the Fourier coefficient taken at the* ***discrete frequency*** *$2\pi n$. We perform sampling in the Fourier space. The Plancherel/Parseval formula becomes:*
>
> $$\|x\|^2 = \sum_n |\langle x, e_n \rangle|^2 = \sum_n |\hat{x}(2\pi n)|^2$$

3.7 The Multi-dimensional Case

Before using the orthonormal basis to expand the RAP triangle analysis, let's take a detour into multi-dimensions to redefine Sobolev regularity. At first glance, transitioning from 1D to an arbitrary dimension seems straightforward; the results seem to translate well. However, there will be a hitch: **approximation results will become very poor**. We'll quantify this with Sobolev.

So, assume we have an orthonormal basis $\{e_n(u), \forall n \in \mathbb{Z}\}$, with $u \in [0,1]$ ([17]), and we would like an orthonormal basis with $u = (u_1, \ldots, u_q) \in [0,1]^q$. The method is simple; we perform a separable product:

Theorem 5 *If $\{e_n(u), \forall n \in \mathbb{Z}\}$ is an orthonormal basis for $L^2([0,1])$, then*

$$\left\{e_n(u) = (e_{n_1}(u_1) \ldots e_{n_q}(u_q)), n = (n_1, \ldots, n_q) \in \mathbb{Z}^q, u = (u_1, \ldots, u_q) \in [0,1]^q\right\}$$

is an orthonormal basis for $L^2([0,1]^q)$.

The proof proceeds in two steps: the first step is to prove that the new vectors (e_n) are orthonormal, and for the second step, we need to show that any function in $L^2([0,1]^q)$ can be decomposed on this basis. For this latter step, we can already notice that it holds for a separable function, meaning a function that can be written as the product $g_1(u_1)g_2(u_2)\ldots g_q(u_q)$. Then, we can show that any function in $L^2([0,1]^q)$ can be approximated by a family of functions constant on small cubes in $[0,1]^q$ (similar to steps in 1D involving steps).

The important thing is that we can extend the Fourier transform to any dimension, because then the orthonormal basis for $L^2([0,1]^q)$ is given by

$$\mathcal{B} = \left\{e_n(u) = e^{i2\pi n.u}, \forall n \in \mathbb{Z}^q, u \in [0,1]^q\right\} \tag{49}$$

with $n.u = \sum_{k=1}^{q} n_k u_k$.

Now, the plan ahead: we can see how the notion of regularity is expressed in q dimensions via Sobolev regularity, which is **equivalent** to the rapid decay of $|\langle x, e_n\rangle|$, providing an **equivalence between regularity and sparsity**. Then, we demonstrate an **equivalence** between **sparse representation** and the quality of **approximation**. We will see the equivalence between the rate of decay of $|\langle x, e_n\rangle|$ and the rate of convergence of the approximation error $\|x - P_{V_M}x\|^2 = \varepsilon_M$ as M tends to infinity. So, we will have equivalences that link the 3 notions of the RAP triangle.

The consequence is that when we tackle neural networks, the object is no longer $x(u)$ but $f(x)$, meaning that the variable is x (an image, for example, depending on u

[17] We can also do the same on $\mathbb{R}$.

with $q = 2$), whose dimensionality explodes (e.g., d the number of pixels, the number of time samples, etc.). Therefore, we will see that the error decay is slow because sparsity is poor, which is due to regularity that is not adapted to the problem (even if we constrained the 100th derivative of f), which is nonlinear.

4. Lecture 27 Jan.

4.1 Exploring the RAP Triangle in a Linear Multidimensional Framework

During this session, we will revisit the RAP (Regularity-Approximation-Sparseness) triangle within a **linear framework** in multiple dimensions. We will begin our exploration by delving into the *regularity* of functions in high dimensions. In doing so, we will rediscover the **Fourier basis** given the linear framework, and establish an *equivalence between regularity and sparsity* in this basis. Next, we will address the *equivalence between sparsity and approximation*. Given that the Fourier representation yields a representation with few coefficients, we will be able to perform *low-dimensional approximations*.

In the linear context, we will also pose the question of what constitutes the "optimal" basis. Here, we will make use of **Principal Component Analysis (PCA)**, namely the **Karhunen-Loève basis**, and we will find Fourier when translation invariance (stationarity) is present. Following this, we will delve into the realm of *non-linearity*. Specifically, we will examine the performance of *single-hidden-layer neural networks*. These networks can be viewed both in a linear context, where we will encounter the **Universal Approximation Theorem**, and in a non-linear context with **Barron spaces**[18]. However, these studies, unfortunately, do not provide answers to algorithm performance in practical scenarios.

A quick reminder: from this point onward, we will primarily use the notation $x(u)$. However, when dealing with high dimensions, we will denote it as $f(x)$. Thus, in low dimensions, u is the underlying variable of a time series or an image, for example. In high dimensions, x becomes the underlying variable in problems of the form $y = f(x)$, and if you consider an image x, its dimensionality is the number of pixels.

[18] Andrew R. Barron, Professor at Yale University

4.2 Regularity of a Function in Multiple Dimensions

The regularity of a function in $L^2([0,1]^q)$ in a dimension $q > 1$ can be examined in a classical manner by looking at partial derivatives. Thus, we aim to control the derivative of $x(u)$ in any direction v, where $\|v\| = 1$:

$$\frac{\partial x(u)}{\partial v} = v \cdot \nabla_u x(u) \tag{50}$$

The partial derivative operator $\partial/\partial v$ is *linear and translation-covariant*, hence it is diagonalizable in a *Fourier basis*. Indeed, by taking an element from the basis (Eq. 48):

$$\frac{\partial e^{i2\pi n.u}}{\partial v} = v \cdot \nabla_u e^{i2\pi n.u} = (i2\pi)\ v \cdot n\ e^{i2\pi n.u} \tag{51}$$

Now, we wish to control any order of derivative of x with respect to any direction v, meaning that it should be square-integrable [19]:

$$\left\| \frac{\partial^p x}{\partial v^p} \right\|_2 = \int_{[0,1]^q} \left| \frac{\partial^p x(u)}{\partial v^p} \right|^2 du \tag{52}$$

and see how this is expressed in the Fourier basis. Let's take the Fourier transform of Eq. 50 ($\omega = 2\pi n$):

$$\widehat{\frac{\partial x}{\partial v}}(2\pi n) = (i2\pi)(v \cdot n)\hat{x}(2\pi n) \tag{53}$$

and, by generalizing through iteration:

$$\widehat{\frac{\partial^p x}{\partial v^p}}(2\pi n) = (i2\pi)^p (v \cdot n)^p \hat{x}(2\pi n) \tag{54}$$

So, the condition of regularity through the control of partial derivatives can be written as:

$$\left\| \frac{\partial^p x}{\partial v^p} \right\|_2 \leq C \overset{Parseval}{\Longleftrightarrow} \sum_{n \in \mathbb{Z}^q} |\hat{x}(2\pi n)|^2 |i2\pi|^{2p} (v \cdot n)^{2p} \leq C \tag{55}$$

The key point here in translating real-domain regularity (derivatives) into the Fourier domain (coefficients) is the fact that the (partial) differentiation operator is diagonal in

[19] $\|z\|_2$ denotes the L^2 norm of z.

the Fourier basis. Reading the condition, this means that when n becomes large, $|\hat{x}(2\pi n)|^2$ must decrease rapidly enough. But this condition must hold for any unit vector $v \in \mathbb{R}^q$. The term $|v \cdot n|^{2p}$ achieves its maximum value when v is collinear with n. Thus, the condition becomes:

$$\sum_{n \in \mathbb{Z}^q} |\hat{x}(2\pi n)|^2 |i2\pi|^{2p} |n|^{2p} \leq C \tag{56}$$

Therefore, the independence of regularity from the direction imposes the equivalence:

$$\boxed{\left\| \frac{\partial^p x}{\partial v^p} \right\|_2 \leq C \Longleftrightarrow \sum_{n \in \mathbb{Z}^q} |\langle x, e_n \rangle|^2 |n|^{2p} \leq C} \tag{57}$$

which is the **Sobolev regularity of degree** p.

4.3 Linear Approximation

4.3.1 Decay of Error and Fourier Coefficients

What is the difference in Equation 57 compared to the one-dimensional case? Essentially, it's the fact that $n \in \mathbb{Z}^q$, but this will have consequences because n lies in a much larger grid. A partial (truncated) sum for the approximation of $x(u)$ in the orthonormal basis can be written as follows:

If we are in one dimension:

$$x_M(u) = \sum_{n=1}^{M} \langle x, e_n \rangle e_n \tag{58}$$

And the error is given by:

$$\|x - x_M\|^2 = \sum_{n > M} |\langle x, e_n \rangle|^2 \tag{59}$$

In arbitrary dimension, the index n is a vector in $\mathbb{Z}^q$. We can restrict its norm, which gives the low-frequency components that select the largest coefficients (Fig. 15). This can be written as:

$$x_M(u) = \sum_{|n| \leq R_M} \langle x, e_n \rangle e_n \tag{60}$$

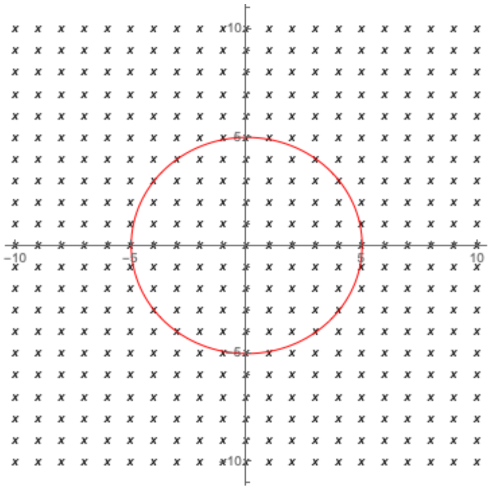

Figure 15 – Low-frequency restriction of the 2D decomposition of $x(u)$.

And the error becomes:

$$\|x - x_M\|^2 = \sum_{|n|>R_M} |\langle x, e_n \rangle|^2 \tag{61}$$

The radius of the sphere, R_M, selects M coefficients. Now, in dimension q, the volume of a ball B_M with radius R_M is given by (assuming q is even for simplification):

$$V(B_M) = \pi^{q/2} R_M^q / (q/2)! \tag{62}$$

Each point on the grid of $\mathbb{Z}^q$ corresponds to a small hypercube and thus, approximately, $V(B_M) \approx M$. Therefore:

$$R_M = M^{1/q} \sqrt{q} \tilde{\gamma}_q (1 + O(\log q/q)) \equiv M^{1/q} \gamma_q \tag{63}$$

(with $\tilde{\gamma}_q = 1/\sqrt{2e\pi} \approx 0.25$)

Now, the idea is to study how the error in Eq. 61 behaves in high dimensions. In fact, the crux of the problem arises from the following observation: when q becomes large, for R_M to become significant and achieve a good approximation, M must grow substantially. The condition:

$$\varepsilon(M) = \|x - x_M\|^2 = \sum_{|n|>M^{1/q}\gamma_q} |\langle x, e_n \rangle|^2 \tag{64}$$

How is this condition expressed in terms of Fourier coefficients (Eq. 57)? To relate the two concepts, there is the following theorem that links the decay of Fourier coefficients and the decay of error:

> **Theorem 6** *Suppose an arbitrary orthonormal basis (not necessarily just the Fourier basis). Then, we have the following equivalence:*
>
> $$\sum_{n\in\mathbb{Z}^q} |\langle x, e_n\rangle|^2 |n|^{2p} \le C \Leftrightarrow \sum_{M=1}^{+\infty} \varepsilon(M) M^{\frac{2p}{q}-1} \le C\gamma'_q \tag{65}$$

What does this mean? If $|\langle x, e_n\rangle|$ decreases rapidly enough, then it is equivalent to saying that $\varepsilon(M) = o(M^{-2p/q})$.

Proof 6. Let's consider the right-hand side term that constrains the rate of decrease of approximation errors by replacing the error with its expression in terms of Fourier coefficients outside the ball:

$$\begin{aligned} A &= \sum_{M=1}^{+\infty} \left(\sum_{|n|>M^{1/q}\gamma_q} |\langle x, e_n\rangle|^2 \right) M^{2p/q-1} \\ &= \sum_{n\in\mathbb{Z}^q} |\langle x, e_n\rangle|^2 \sum_{M=1}^{|n|^q\gamma_q^{-q}} M^{2p/q-1} \sim \sum_{n\in\mathbb{Z}^q} |\langle x, e_n\rangle|^2 \frac{|n|^{2p}\gamma_q^{-2p}}{2p/q} \\ &\sim \gamma'_q \times \sum_{n\in\mathbb{Z}^q} |\langle x, e_n\rangle|^2 |n|^{2p} \end{aligned}$$

We used:

$$\int_1^{a+1} u^s du \le \sum_{M=1}^{a} M^s \le \int_0^a u^s du = \frac{a^{s+1}}{s+1}$$

which gives an equivalent of the sum when a tends to infinity (and $s < 0$). So, A as the right-hand term of the equivalence is proportional to the left-hand term, which indeed establishes the relationship between the decay of error and Fourier coefficients. ■

This is a very strong result because from the rate of error decay, we can deduce the form of function regularity. For example, if we know that the function is twice differentiable but not three times, then the error decay is true for $p = 2$ but not for $p = 3$. Once

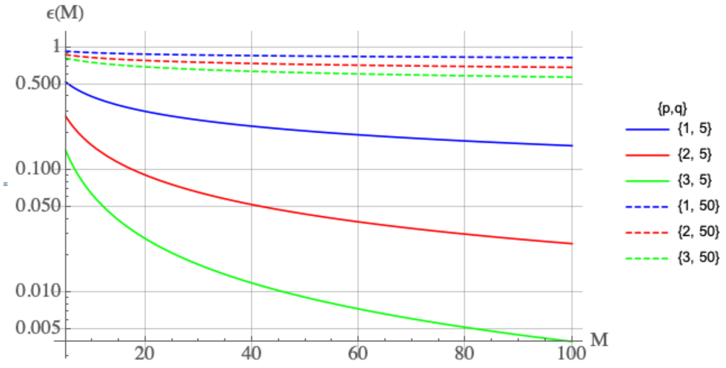

FIGURE 16 – Examples of $\varepsilon(M)$ decay depending on the values of p (derivative order) and q (dimension of space).

again, the decay exactly reflects the order of regularity. This result is extensively used in Approximation Theory to demonstrate equivalences between Sobolev-type regularity and linear approximation forms, whether in Fourier bases or equivalent bases.

4.3.2 Curse of Dimensionality

As previously noted, the error decay satisfies the fact that

$$\varepsilon(M) = o(M^{-2p/q}) \tag{66}$$

and this is illustrated in Figure 16. For example, if we want to achieve an error of ε, then the number of coefficients M must roughly satisfy

$$M \sim \varepsilon^{-q/2p} \tag{67}$$

which means that **this number must increase exponentially with dimensionality** q. This is the core problem because even if the function is, for instance, continuously differentiable 100 times, in high dimensions, M will explode (considering x as the variable of $f(x)$ in which case $q \sim 10^{4-6}$). **Thus, as the dimension significantly increases, in the linear framework, we quickly become limited**. However, it's true that if we're dealing with time series where $q = 1$, there's no issue, for images with $q = 2$, it works but less effectively.

4.3.3 Low-Frequency Filter

We obtain a linear approximation x_M by taking a partial series of the signal $x(u)$:

$$x_M(u) = \sum_{|n| \leq R_M} \langle x, e_n \rangle e_n \tag{68}$$

The Fourier coefficients of this approximation are those of x, but only at low frequencies. That is, we have the simple relationship:

$$\widehat{x_M}(2\pi n) = \hat{x}(2\pi n) \ \mathbb{1}_{|n| \leq R_M} \tag{69}$$

Now, the operator $\mathbb{1}_{|n| \leq R_M}$ takes values of either 1 or 0. It's diagonal in Fourier, so it's a convolution. In fact, as in the 1D case, we can view x_M **as the convolution of** x **with a Dirichlet kernel or low-pass filter** h_M **that equals 1 inside the ball** B_M:

$$x_M = x * h_M \tag{70}$$

Thus, we rediscover the 1D idea that if a function is regular, we eliminate high frequencies to obtain a linear approximation. This result is used in single-hidden-layer neural networks.

4.4 Discovering the Right Basis: Unsupervised Learning

When we start with data, we generally don't have any *a priori* knowledge of the regularity of the underlying function. In the absence of regularity, we enter the RAP triangle through the approximation-sparsity side. Thus, the question arises: given the data alone, what is the linear approximation x_M that minimizes the error?

Speaking of linear approximation, it involves projecting x onto a linear space V_M (see Fig. 8) to obtain x_M. So, the question can be reformulated as: what is the linear space V_M that allows us to approximate signals x to minimize $\varepsilon(M)$? The first observation is that if we have only one signal x, we can choose any hyperplane containing x, which trivially gives $\varepsilon(M) = 0$. Therefore, we need several signals $\{x_i\}_{i \leq N}$ from a space Ω, and we will use **Unsupervised Learning** to obtain the best approximation space V_M.

Initially, the "generic" signal $x(u)$ is of finite dimension $\mathbb{R}^d$, where u takes d values. However, x is a *random vector* in $\mathbb{R}^d$. And we need to establish the measure of the approximation error, which must be minimized for all $x \in \Omega$ to obtain the optimal V_M. Thus, we would like to obtain:

$$\underset{x_M \in V_M}{\text{Min}} \ E(\|x - x_M\|^2) \tag{71}$$

A fundamental concept when using **quadratic error** is that minimization will depend only on one thing: the **covariance**. Let's have a quick reminder to fix the notations:

$$E[x] = \mu \in \mathbb{R}^d, \qquad E[(x(u) - E[x(u)])(x(u') - E[x(u')])^*] = \mathbf{K}(u, u') \in \mathbb{R}^{d \times d} \tag{72}$$

and $\mathbf{K}$ is a positive definite Hermitian matrix.

How is the covariance matrix $\mathbf{K}$ related to the approximation error with a random vector? In fact, $\mathbf{K}$ will completely characterize linear combinations. A linear combination is written as the dot product of x with $z \in \mathbb{R}^d$, a deterministic vector (i.e., fixed with respect to the randomness of x):

$$\langle x, z \rangle = \sum_{u=1}^{d} x(u) z(u) \tag{73}$$

If we consider another linear combination $\langle x, z' \rangle$ and want to correlate it with the previous one, then (let's simplify the notations by assuming that $E(x) = 0$, so we consider the vector x from which we subtract the mean μ):

$$\begin{aligned} E[\langle x, z \rangle \langle x, z' \rangle^*] &= E[\sum_u x(u) z^*(u) \times \sum_{u'} x^*(u') z'(u')] \\ &= \sum_{u,u'} z^*(u) z'(u') E[x(u) x^*(u')] \\ &= \sum_{u,u'} z^*(u) z'(u') \mathbf{K}(u, u') \\ &= \langle K z', z \rangle = z^T . \mathbf{K} z' \end{aligned} \tag{74}$$

So, the entire characterization of the correlation is given by the matrix $\mathbf{K}$. Note that if

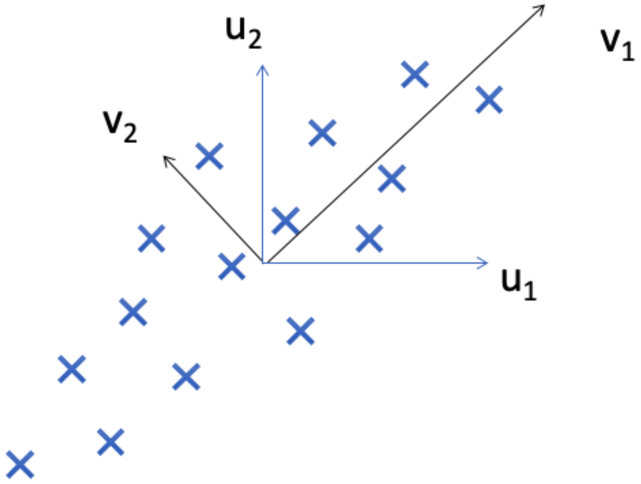

FIGURE 17 – Illustration of the principal axes of the covariance matrix.

$z = z'$:

$$E[|\langle x, z\rangle|^2] = \langle Kz, z\rangle \tag{75}$$

Now, let's consider the (expected) error of the approximation:

$$E(\|x - x_M\|^2) = E(\sum_{|n|>R_M} |\langle x, e_n\rangle|^2) = \sum_{|n|>R_M} \langle Ke_n, e_n\rangle \tag{76}$$

We notice that once the basis is fixed, **the error depends only on the covariance matrix**. The best basis will depend on the properties of this covariance matrix. Note that a geometric interpretation of **K** is given by **the principal axes** of a population of vectors x (Fig. 17). The axis with the greatest variability corresponds to the eigenvector with the largest eigenvalue.

Here's a widely-used theorem in data analysis and functional analysis because it tells you why the Fourier basis is often optimal:

> **Theorem 7** *For any M, $E(\|x - x_M\|^2 = \sum_{|n|>R_M} E(|\langle x, e_n\rangle|^2)$ is minimized if the elements of the basis $\{e_n\}_{n\leq d}$ diagonalize the covariance matrix* **K** *associated with*

> *x, with eigenvalues*
>
> $$\lambda_n = \langle K e_n, e_n \rangle \geq 0$$
>
> *that are ordered in decreasing order. Another way to express the same thing is that*
>
> $$E(\|x - P_{V_M} x\|^2)$$
>
> *is minimized if V_M is generated by the first M eigenvectors of the covariance matrix* $\mathbf{K}$.

Proof 7. To begin the proof, let's apply the Pythagorean Theorem, knowing that x_M is the orthogonal projection of x onto the hyperplane V_M:

$$E(\|x - x_M\|^2) = E(\|x\|^2) - E(\|x_M\|^2)$$

In this context, we aim to minimize the approximation error concerning all possible x_M, which is equivalent to maximizing $E(\|x_M\|^2)$. Now,

$$E(\|x_M\|^2) = \sum_{|n| \leq R_M} E(|\langle x, e_n \rangle|^2)$$

In fact, the condition $|n| \leq R_M$ defines the M vectors of the basis $\{e_n\}_{n \leq d}$, so we can rewrite the sum as $\sum_{n=1}^{M}$. Let's consider the basis that diagonalizes the matrix $\mathbf{K}$ with (λ_n) as the spectrum of eigenvalues, denoted as $\{\bar{e}_n\}_{n \leq d}$. This is the **Karhunen-Loève basis** [20], defined such that

$$\langle K \bar{e}_n, \bar{e}_n \rangle = \lambda_n \geq 0 \quad \text{(decreasing)} \qquad\qquad \langle K \bar{e}_n, \bar{e}_{n'} \rangle = 0 \quad (n \neq n')$$

So, the idea is to show that $E(\|x_M\|^2)$ is larger in the Karhunen-Loève basis, so we need to express e_n in this particular basis:

$$e_n = \sum_{k=1}^{d} \langle e_n, \bar{e}_k \rangle \bar{e}_k$$

[20] named after Kari Karhunen (1915-92), a Finnish mathematician, and Michel Loève (1907-79), a Franco-American mathematician.

Consequently, it follows that

$$\begin{aligned}
E(|\langle x, e_n\rangle|^2) &= E(|\sum_{k=1}^{d}\langle e_n, \bar{e}_k\rangle\langle x, \bar{e}_k\rangle|^2) \\
&= \sum_{k,k'}\langle e_n, \bar{e}_k\rangle\langle e_n, \bar{e}_{k'}\rangle^* E(\langle x, \bar{e}_k\rangle\langle x, \bar{e}_{k'}\rangle^*) = \sum_{k,k'}\langle e_n, \bar{e}_k\rangle\langle e_n, \bar{e}_{k'}\rangle^*\langle K\bar{e}_k, \bar{e}_{k'}\rangle \\
&= \sum_k \lambda_k|\langle e_n, \bar{e}_k\rangle|^2
\end{aligned}$$

So, the expression for $E(\|x_M\|^2)$ becomes

$$E(\|x_M\|^2) = \sum_{n=1}^{M}\sum_{k=1}^{d}\lambda_k|\langle e_n, \bar{e}_k\rangle|^2 = \sum_{k=1}^{d}\lambda_k\left(\sum_{n=1}^{M}|\langle e_n, \bar{e}_k\rangle|^2\right)$$

and we aim to maximize this expression by choosing the basis $\{e_n\}_{n\leq M}$ of V_M (considering the degrees of freedom of the problem). Now,

$$0 \leq c_k = \sum_{n=1}^{M}|\langle e_n, \bar{e}_k\rangle|^2 \leq \sum_{n=1}^{d}|\langle e_n, \bar{e}_k\rangle|^2 = \|\bar{e}_k\|^2 = 1$$

and, on the other hand,

$$\sum_{k=1}^{d} c_k = \sum_{n=1}^{M}\sum_{k=1}^{d}|\langle e_n, \bar{e}_k\rangle|^2 = \sum_{n=1}^{M}\|e_n\|^2 = M$$

So, the $\{c_k\}_{k\leq d}$ are numbers in the range $[0,1]$ whose sum is equal to M. Now, the $\{\lambda_k\}_{k\leq d}$ are either positive or zero and arranged in decreasing order. Hence, to maximize $E(\|x_M\|^2)$, it suffices to set the first M c_k values to 1 and the rest to 0:

$$c_k = \begin{cases} 1 & k \leq M \\ 0 & k = M+1, \ldots, d \end{cases}$$

Thus, for all $k \leq M$, $\sum_{n=1}^{M}|\langle e_n, \bar{e}_k\rangle|^2 = 1$, so the orthonormal vectors $\{e_n\}_{n\leq M}$ are in the space spanned by the first M vectors of the Karhunen-Loève basis, and finally, the same applies to V_M. ■

This theorem is very important in practice, of course, and also because we realize that the approximation problem becomes an operator diagonalization problem. Moreover, in practice, the concept of stationary signals is often encountered, and we will see that one can guess the orthonormal basis then.

4.5 Stationary Signals

Consider Second-Order Stationary Random Processes[21], and let $p(x)$ be the probability density of $x \in \mathbb{R}^d$. What happens to the probability density if the signal is translated:

$$x_\tau(u) = x(u-\tau) \Rightarrow p(x_\tau)? \tag{77}$$

In this context, we say the **process is stationary**[22] if

$$\forall k, \alpha,\ p(x(u_1), x(u_2), \ldots, x(u_k)) = p(x(u_1-\alpha), x(u_2-\alpha), \ldots, x(u_k-\alpha)) \tag{78}$$

As a consequence, the mean of the signal is a constant. This is the case, for example, in imaging or audio recording when there is no reference point. Additionally,

$$E[f(x(u_1), x(u_1+\tau))] = E[f(x(u_1-\alpha), x(u_1-\alpha+\tau))] = E[f(x(0), x(\tau))]$$

which is a function of τ alone. This implies that if we take $f(x,y) = xy-\mu$, the covariance matrix has the following property:

$$\forall u, u',\ K(u,u') = K(u-u') \tag{79}$$

So, when we apply the operator K of a stationary process to a function g, for example, we have

$$K.g(u) = \sum_{u'} g(u')K(u,u') = \sum_{u'} g(u')K(u-u') = (K*g)(u) \tag{80}$$

This is a **convolution operation** by K.

Now, a convolution operator is **diagonalized in a Fourier basis**. So, as soon as we

[21] We weaken the notion of stationarity for ease of study.

[22] a strong assumption

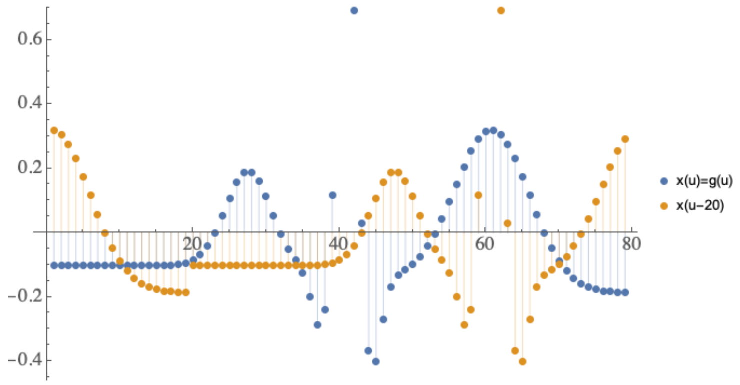

FIGURE 18 – Example of $g(u)$ and a periodicized translated version (note: the mean of g is zero).

have a **stationary process, the Karhunen-Loève basis is the Fourier basis**, namely

$$\left\{\bar{e}_k(u) = e^{i2\pi ku/d}\right\}_{1\leq u\leq d} \tag{81}$$

if we are in a discrete space with d values where we use modulo d periodicity of u to define translation (note: the frequencies are $2\pi k/d$).

Returning to the notion of uniform regularity, if, for example, the derivative of the signal is bounded, the same holds for the translated signal, defining classes of functions that are translation-invariant. **And as the optimal Karhunen-Loève basis is identical to the Fourier basis in this case, we can never do better in the context of linear approximation.**

So, we have achieved the optimum in what we can do linearly, but is it satisfactory? Let's take a function $g(u)$ with $u \in [1, d]$, and any signal x is defined from g as follows (v.a: random variable) (Fig. 18):

$$x_\tau(u) = g((u-\tau)\ \mathrm{mod}[d]), \quad \tau \in \{1,\dots,d\} \text{ uniform random variable} \tag{82}$$

The signal x is a random process, and since τ has a uniform distribution, we are equally likely to "see" x or x_τ, so the process is stationary. What is the matrix $\mathbf{K}$? First of all,

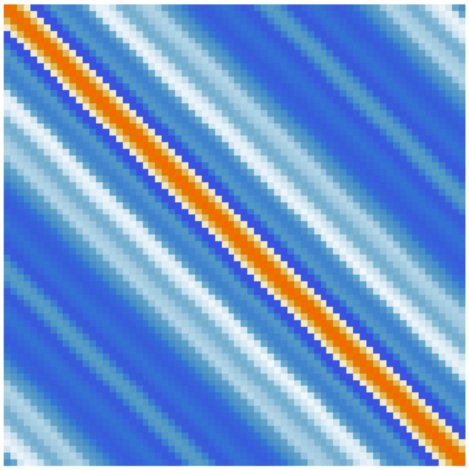

FIGURE 19 – The matrix $K(u, u')$ of the signal (Fig. 18), which clearly has a banded structure indicating that the matrix depends only on the variable $u - u'$.

note that the expectation of x is simply the mean of g (invariant under translation):

$$E[x_\tau(u)] = \sum_{\tau=1}^{d} p(\tau) g((u-\tau)\text{mod}[d]) = \frac{1}{d}\sum_{\tau=1}^{d} g((u-\tau)\text{mod}[d]) = \frac{1}{d}\sum_{u'=1}^{d} g(u') \qquad (83)$$

Let's assume this mean is zero to simplify calculations, then it follows:

$$\begin{aligned} E[x_\tau(u)x_\tau(u')] &= \sum_{\tau=1}^{d} p(\tau)\ g((u-\tau)\text{mod}[d])\ g((u'-\tau)\text{mod}[d]) \\ &= \frac{1}{d}\sum_{\tau=1}^{d} g((u-\tau)\text{mod}[d])\ g((u'-\tau)\text{mod}[d]) \\ &= \frac{1}{d}\sum_{u''=1}^{d} g(u'')g(u''-(u-u')))\text{mod}[d]) = K(u-u') \qquad (84) \end{aligned}$$

So, we indeed have a stationary signal (Fig. 19) with a covariance matrix that depends only on the variable $u - u'$. Thus, through Theorem 7, we arrive at the Fourier basis for performing low-frequency linear approximation. However, the signal $g(u)$ can have discontinuities, so it's not always legitimate to proceed with a low-frequency approximation of it. We want to go for a nonlinear approximation that adapts the sampling to the regularity of $g(u)$. From the perspective of the optimal basis, we won't necessarily take the first M vectors, but we'll need to cleverly choose the vectors. What we will see then is that Theorem 7 is no longer valid. We will discover the need to redefine other classes of regularity, which are certainly more complex but worth it for certain types of problems. We will also see that if we have a purely "linear" view when analyzing single-layer neural networks, everything seems simple, and we don't do significantly better than the Fourier basis (or PCA). However, as soon as we take a "non-linear" perspective, the theorems become more complicated to interpret.

5. Lecture 3 Feb.

During this session, we will explore the RAP triangle from a *non-linear* perspective. It's worth recalling that in a *linear* context, obtaining a low-dimensional approximation at the heart of data analysis is equivalent to adopting a sparse representation that concentrates information on a few coefficients. This notion of sparsity is equivalent to the regularity pattern of the underlying function. In the linear case, for the class of functions invariant under translation, we've seen that the optimal basis is the Fourier basis. In a more general context, we've also seen that the best basis is the one that diagonalizes the covariance matrix, namely the Karhunen-Loève basis, which coincides with the Fourier basis for stationary processes.

The question that arises now is whether we can do better by taking a *non-linear* perspective? At the end of the previous section, we pointed out that an improvement might be conceivable in cases where the underlying function exhibits discontinuities, meaning it's only *piecewise regular*, as exemplified by the function in Figure 10. A linear approximation would involve considering the function as *uniformly regular*, implying regular sampling and the exclusion of high frequencies. This approach, however, limits the quality of approximation around potential discontinuities. In such cases, adaptation on a case-by-case basis becomes necessary. But how can we do this generically or automatically? When considering *adaptive sampling*, we assume that the regularity of the function is not uniform everywhere, implying potential singularities. Note that these discontinuities/singularities contain crucial information, such as object boundaries in 2D or musical note attacks. Hence, the structure of signals resides in the *high-frequency components*. The question then becomes: can we capture these discontinuities by having a sparse representation that provides a lower-dimensional approximation of higher quality than that obtained in the linear context?

5.1 Non-linear Sparse Representation

We position ourselves in the context of searching for an orthonormal basis, meaning the signal x decomposes as follows:

$$x = \sum_{n=1}^{\infty} \langle x, e_n \rangle e_n \tag{85}$$

In the **linear** case, the approximation is performed by considering the truncated partial sum with the first M **coefficients**. In Fourier, this corresponds to the low frequencies. In the **non-linear** case, we'll proceed with **choosing the M coefficients**, and thus we can write:

$$x_M = \sum_{n \in I_M} \langle x, e_n \rangle e_n \tag{86}$$

where I_M is a set of indices n depending on x (while keeping $|I_M| = M$). Since we have an orthonormal basis, the error is straightforward to formulate as:

$$\varepsilon_M = \|x - x_M\|^2 = \| \sum_{n \notin I_M} \langle x, e_n \rangle e_n \|^2 = \sum_{n \notin I_M} |\langle x, e_n \rangle|^2 \tag{87}$$

The choice of I_M is made with the aim of minimizing this approximation error, and this naturally leads to the definition:

$$I_M = \{ n \ / \ |\langle x, e_n \rangle| \geq T_M \} , \qquad \text{where } \mathrm{T_M} \text{ such that } |I_M| = M \tag{88}$$

which allows us to write:

$$\varepsilon_M = \sum_{|\langle x, e_n \rangle| < T_M} |\langle x, e_n \rangle|^2 \tag{89}$$

The question is whether ε_M calculated in this way is significantly smaller than what we would have obtained with a linear approach. We will see that the answer is "yes" if the basis provides a sparse representation. Therefore, Approximation and Sparsity are once again two intimately related concepts.

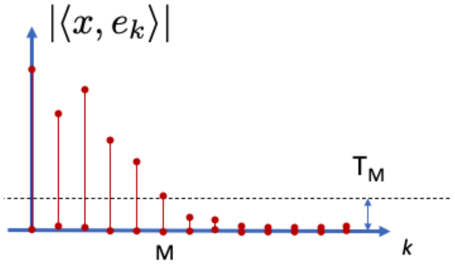

FIGURE 20 – Selection of the first M inner products once ordered in decreasing order, defining the threshold T_M.

Sparsity is manifested by the decreasing inner products but once ordered in decreasing order:

$$|\langle x, e_k \rangle| \geq |\langle x, e_{k+1} \rangle| \tag{90}$$

As illustrated in Figure 20, the error corresponds to the sum of the squares of terms where $k > M$ or whose intensity is below the threshold T_M. Thus,

$$\varepsilon_M = \sum_{k=M+1}^{\infty} |\langle x, e_k \rangle|^2 \tag{91}$$

That being said, sparsity in the non-linear sense can also be represented without sorting coefficients in descending order, as shown in Figure 21, where **the coefficients exceeding the threshold T_M are few in number** and are **located anywhere** depending on the signal x. Furthermore, the residual error decreases towards 0 as the threshold is lowered. How fast does it decrease?

5.1.1 Rate of Decrease of Non-linear Error

As in the linear case (Th. 6), we can constrain the rate of decrease of the error with the following theorem:

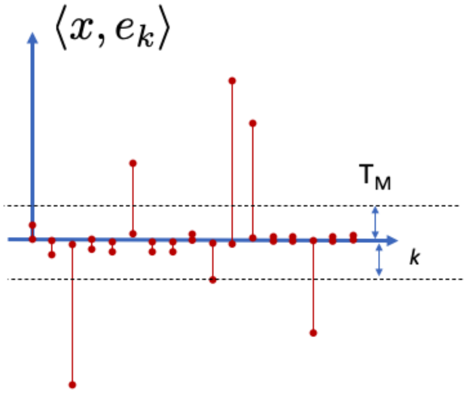

FIGURE 21 – Selection through the threshold T_M of non-ordered inner products.

Theorem 8 *If we order the inner products in decreasing order, and p is a derivative order, if*

$$\sum_{k=1}^{\infty} |\langle x, e_k \rangle|^2 k^{2p} < \infty \tag{92}$$

which is equivalent to $|\langle x, e_k \rangle|^2 = o(k^{-2p-1})$ *and characterizes the rate of decrease of inner products, then*

$$\sum_{M=1}^{\infty} \varepsilon_M M^{2p-1} < \infty \tag{93}$$

which means that $\varepsilon_M = o(M^{-2p})$. *In fact, there is an equivalence between the two rates of decrease.*

In the linear case, we have the index $k \in \mathbb{Z}^q$, which gives $\varepsilon_M = o(M^{-2p/q})$. In the case of the theorem above, $k \in \mathbb{Z}$, otherwise, the proof is identical. So, we have obtained an equivalence between the concepts of low-dimensional approximation and sparse representation of the RAP triangle. However, even though there are similarities between the *linear* and *non-linear* cases, for the latter, **the result is obtained by ordering the inner products by importance, which incidentally depends on the signal** x. Nevertheless, it would be desirable to obtain a constraint independent of the order of coefficients since that would be more practical.

5.1.2 Sparsity and ℓ^α Norm

There is an important theorem that lies at the heart of Approximation mathematics and forms the basis for many algorithms. It states that the rate of decrease of ordered coefficients is characterized by ℓ^α norms. Instead of dealing with the L^2 norm in an orthonormal basis, for example via

$$\sum |\langle x, e_n\rangle|^2 = \|x\|^2$$

we will focus on norms defined as

$$\left(\sum |\langle x, e_n\rangle|^\alpha\right)^{1/\alpha}$$

especially with $\alpha < 2$.

Theorem 9 *If for $\alpha < 2$*

$$C_\alpha = \left(\sum_{q=1}^{\infty} |\langle x, e_q\rangle|^\alpha\right)^{1/\alpha} < \infty \tag{94}$$

then, by ordering the inner products, the one of rank k satisfies a decreasing constraint, resulting in a rate of decrease of the error:

$$|\langle x, e_k\rangle| \leq C_\alpha\ k^{-1/\alpha} \qquad \text{and} \qquad \varepsilon_M \leq \frac{C_\alpha^2}{\frac{2}{\alpha}-1} M^{-\left(\frac{2}{\alpha}-1\right)} \tag{95}$$

It is important to note that ***the ℓ^α norm of x, i.e., C_α, controls both the rate of decrease of ordered inner products and that of the error ε_M.***

Proof 9. We start from the definition of C_α with an orthonormal basis $\{e_n\}$ and a signal x:

$$C_\alpha^\alpha = \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^\alpha$$

then we rearrange the indices with a descending order of inner products, which introduces a new set of indices n_q:

$$C_\alpha^\alpha = \sum_{q=1}^{\infty} |\langle x, e_{n_q} \rangle|^\alpha$$

Now, we can truncate the series to the first k terms to lower bound the series:

$$C_\alpha^\alpha \geq \sum_{q=1}^{k} |\langle x, e_{n_q} \rangle|^\alpha \geq k |\langle x, e_{n_k} \rangle|^\alpha$$

So, in other words, the coefficient of rank k satisfies the constraint:

$$|\langle x, e_{n_k} \rangle| \leq C_\alpha k^{-1/\alpha}$$

which is what we wanted to prove[23]. ■

This theorem indicates that the smaller α is, the faster the inner products and the error decrease. Therefore, we would like to consider ℓ^α norms with α as small as possible.

Now, as long as $\alpha \geq 1$, C_α is a *convex function* of the inner products, and this is no longer the case when $\alpha < 1$. Why is this important? The reason is that if we do not know the basis $\{e_n\}$, we need to discover it. In this case, we will want to minimize the ℓ^α norm, which becomes a sort of cost function. In this context, having a convex function allows the use of optimization algorithms. Therefore, in practice, we work with α in the interval $[1, 2)$, as small as possible, hence **the use of the ℓ^1 norm** in optimization problems while ensuring sparsity.

5.2 Application to Single Hidden Layer Neural Networks

In this section, we will put into practice everything we have learned so far to study the classification/regression problem of the form $f(x) = y$ using a single hidden layer

[23] Note that to obtain the version with "little o", S. Mallat suggests taking the upper bound not from the sum from $q = 1$ to k but from $k/2$ to k, which tends to 0.

neural network. We will encounter two pitfalls that will lead us to radically change our perspective.

For this type of problem, let's recall that our goal is to approximate the function f, which maps x to y, where y is either an integer type for classification or a real type for regression. Here, x is the variable, and its domain is either $\mathbb{R}^d$ or $[0,1]^d$ for bounded signals, but most importantly, with $d \approx 10^{4-6}$. So far, our object of study has been $x(u)$ with u in very low dimension q ($u \in [0,1]^q$). We know that when q starts to increase significantly, we encounter the curse of dimensionality [24].

Each neuron m in a network with M hidden neurons (Fig. 5) first calculates a dot product with a vector w_m as follows:

$$x.w_m = \sum_{u=1}^{d} x(u) w_m(u) \tag{96}$$

Then, we apply a non-linearity ρ (ReLU, sigmoid, hyperbolic tangent, cosine, etc.) and a bias b_m:

$$\rho(x.w_m + b_m) \in \mathbb{R} \tag{97}$$

Finally, we linearly combine the M non-linearities to obtain $\tilde{f}$:

$$\tilde{f}(x) = \sum_{m=1}^{M} \alpha_m \rho(x.w_m + b_m) \tag{98}$$

In this case, we choose (this is an *a priori* assumption, a topic from the 2020 lecture) to decompose $\tilde{f}$ into a family of functions $\{\rho(x.w_m + b_m)\}_{m \leq M}$. We perform **a projection of f into the space generated by this family of functions**. The immediate question is: what is the size of the error made?

To answer this question, we will consider the mean squared error, for example, if the signals are in $[0,1]^d$ (e.g., pixel values of an image):

$$\|f - \tilde{f}\|^2 = \int_{x \in [0,1]^d} |f(x) - \tilde{f}(x)|^2 \, dx \tag{99}$$

The idea is to show that by using what we have learned so far, we will be able to provide

[24] For example, refer to the 2018 and 2019 lectures.

the "usual" answers on the subject.

5.2.1 Universal Approximation (Linear Perspective)

One initial result[25] is the theorem of universal approximation, which states that, using non-polynomial nonlinearities, the approximation error tends to zero as M approaches infinity. This theorem was demonstrated and refined mainly from 1988-92 by the community of mathematicians specializing in approximation[26] using well-known techniques, as it is primarily a result of linear approximation. So, where does the fact that the error tends to zero come from?

To shed light on this, let's take the viewpoint of *linear approximation*. We ask whether there exists a family of generic functions that will be sufficient to approximate any function f as the number of components (M) tends to infinity. But, as we've seen, *a family of functions in linear approximation implies Fourier*. So, the question is how to construct a Fourier basis from the family $\{\rho(x.w_m + b_m)\}$? Initially, let's take the nonlinearity as the function $\rho(a) = e^{ia}$. Then, we have:

$$\tilde{f}(x) = \sum_{m=1}^{M} \alpha_m e^{i(x.w_m+b_m)} = \sum_{m=1}^{M} \alpha_m e^{i\,b_m}\ e^{i\,x.w_m} \tag{100}$$

which resembles a Fourier series. Now, we need to learn the weights w_m (from the network's perspective), but in Fourier analysis, these correspond to the frequencies of the decomposition. So, as long as the underlying function f has some *smoothness*, which is assumed by all theorems dealing with this linear perspective, we focus on a *low-frequency approximation* (Sec. 4.3). If we consider the case where $x \in [0,1]^d$, then $w_n = 2\pi n$ with $n \in \mathbb{Z}^d$ (Sec. 3.7). Thus, we get:

$$\tilde{f}(x) = \sum_{m=1}^{M} \alpha_m e^{i\,b_m}\ e^{i2\pi\,x.m} \tag{101}$$

And since we have an orthonormal Fourier basis, we have:

$$\alpha_m e^{i\,b_m} = \langle f(x), e^{i2\pi\,x.m} \rangle = \int_{x\in[0,1]^d} f(x) e^{-i2\pi\,x.m}\ dx = \hat{f}(2\pi m) \tag{102}$$

[25] See the 2019 course for a version of the proof.

[26] Notable contributors include George Cybenko, Kurt Hornik, Allan Pinkus, etc.

Furthermore, we have the guarantee of obtaining the minimum (linear) approximation since the Fourier basis is optimal in this case. We know that f is real [27],so $\hat{f}^*(\omega) = \hat{f}(-\omega)$, and

$$\tilde{f}(x) = \sum_{m=1}^{M} \alpha_m \cos(w_m.x + b_m) \tag{103}$$

Now, we know that if $f \in L^2([0,1]^d)$, meaning that f has finite energy, which is a reasonable assumption, a result on Fourier series tells us that (unless something extraordinary happens):

$$\lim_{M\to\infty} |f(x) - \sum_{m=1}^{M} \alpha_m \cos(w_m.x + b_m)|^2 = 0 \tag{104}$$

So, the theorem of universal approximation is roughly based on the notions we have seen previously.

The "roughly" means, for instance, that the theorem works for other types of non-linearities than non-polynomial exponentials. Why? We'd like to *change the basis in 1D* between the families $\{\rho(w.x+b)\}_{w,b}$ and $\{\cos(w.x+b)\}_{w,b}$ with ρ being ReLU, sigmoid, etc. For example, between ReLU and cosine, it suffices to show that cosine can be approximated by piecewise linear functions on sufficiently fine steps (e.g., see 2019 Course Sec. 5.3.2). Additionally, the aforementioned convergence is in terms of the L^2 norm, whereas the theorems achieve uniform convergence (i.e., in the "sup" norm on the space in which x evolves). However, these are refinements on **a strong dominance of linear approximation and the Fourier basis**.

Nevertheless, having used the Fourier basis tells us that if we want rapid convergence, we must impose regularity constraints on the class of functions f. For example, if f belongs to a Sobolev space of degree p (Sec. 3.5), meaning that all derivatives of order p have finite energy, then (Th. 6):

$$\|\tilde{f}_M - f\| = o(M^{\frac{-2p}{d}}) \tag{105}$$

But let's examine this result more closely: it tells us that even if the functions are very smooth (p large), which tends to speed up convergence, the fact that d can be very large

[27] Let's assume we are in this general case, which is not restrictive for practical cases like sound processing or image analysis.

($d \approx 10^{4-6}$) significantly reduces the rate of convergence unless we exponentially increase the number of hidden neurons (M). **We find ourselves facing the curse of dimensionality, and most importantly, we fail to understand, in this framework, the quite remarkable results of neural networks.** But that's not the end of the story...

5.2.2 The Non-linear Perspective

What was noticed in 1992 (e.g., A. Barron) is that the problem of finding an approximation $\tilde{f}$ should not be considered for a class of functions with a generic basis. Instead, it's necessary to specialize, not to say **tailor the reasoning to the specific function f that interests us for the posed problem**[28]. Therefore, searching for the general case is in vain, as the specific case is the only one that interests us, at least initially, especially from a "non-linear" viewpoint.

We need to **adapt the projection basis**; thus, in this context, we need to adapt the (w_n, b_n) with respect to the object to approximate, namely, f. In the previous course, we found a way to adapt the decomposition to x by selecting the inner products $\langle x, e_n \rangle$ beyond a threshold $T_M(x)$ chosen to have only M components. What should we do now in our problem $f(x) = y$?

We will use the relationship between sparsity (low-dimensional approximation) and ℓ^α norms (Sec. 5.1.2). Because what do we need to **overcome the curse of dimensionality**? We need to find a basis in which the inner products $|\langle f, e_k \rangle|$ decrease in an orderly manner as $1/k^\alpha$ (or even $\alpha = 1$ for convexity) because then, according to Theorem 9, we can deduce that the error will be constrained by:

$$\|f - \tilde{f}_M\|^2 \leq \frac{C_\alpha^2(f)}{\frac{2}{\alpha} - 1} M^{-\left(\frac{2}{\alpha}-1\right)} \tag{106}$$

meaning that **the convergence speed will no longer depend on the dimension d** of the variable x. Thus, to make this result work, we need **the ℓ^α norm of f to be bounded**.

[28] NDJE: What will be the problem later is to explain why weights learned to recognize cats/dogs are entirely relevant for recognizing boats/cars, i.e., explaining a form of generality of the functions learned by convolutional neural networks.

The first article that put this scheme into practice dates back to 1993 (A. Barron [29]). Its result can be formulated as follows. First, note that the basis $\{e_n\}$ chosen by A. Barron is the Fourier basis with $e_n = e^{i2\pi n}$. Second, instead of considering functions with Sobolev regularity, for which we would be subject to the curse of dimensionality, let's consider **a new type of regularity (*Barron spaces*)**.

> **Theorem 10** ***(A Barron 1993)***
> *If the ℓ^α norm of f is finite and take $\alpha = 1$, and if we consider the Fourier basis ($e_n = e^{i2\pi n}$) such that*
> $$\sum_n |\langle f, e_n\rangle| \leq C_f \tag{107}$$
> *then the neural network with M neurons can provide an approximation $\tilde{f}_M$ such that:*
> $$\|f - \tilde{f}_M\|^2 \leq \frac{C_f^2}{M} \tag{108}$$

Notice that **the dimension d no longer appears**, and we have breached the curse of dimensionality. Later on, we were able to manipulate the parameter α ($f \in \mathcal{B}^\alpha$) and derive results that we understand well, given the concepts covered in previous courses. **So, the curse of dimensionality is gone, but does this explain the results of neural networks (convolutional)?**

The underlying problem we face with this result is the following: does the class of functions $f \in \mathcal{B}^\alpha$ reflect the class of functions we encounter in practice and that neural networks approximate well? **In other words, the result holds for a class of functions, but is this class representative of the functions we are practically dealing with (image classifications, sound analysis, text analysis, regression, etc.), and do neural networks approximate them very well?** Unfortunately, **the answer is negative**, and the community has realized this quite clearly, to the extent that there is a clear divergence between this type of mathematical theorem and what is practiced every day by those who use/implement neural networks. So, what does this mean? Be aware that Barron's theo-

[29] https://www.researchgate.net/publication/3078296_Barron_AE_Universal_approximation_bounds_for_superpositions_of_a_sigmoidal_function_IEEE_Trans_on_Information_Theory_39_930-945

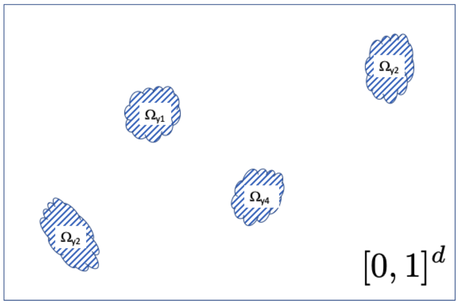

FIGURE 22 – The volume of object classes Ω_y is tiny compared to the space in which x operates.

rem is correct, as are those who followed in his footsteps. There's no doubt about that. However, it's essential to understand why this is not (yet) the right approach.

5.2.3 A New Perspective: The Bayesian Approach

Let's try to understand why the Barron-like approach doesn't exhaust the subject. Take the case of image classification (dog, cat, boat, coffee machine...), but the problem is the same with sounds. In 99.9% of cases (not 100% to account for some human error), there is no ambiguity when we take an image. So, we associate class indices y with different x (images), and we attempt to solve $y = f(x)$. We can represent the members of class y as:

$$\Omega_y = \{x \mid f(x) = y\} \tag{109}$$

When we change y, we realize that the volume of Ω_y is tiny compared to the size of the space in which x operates (Fig. 22). In other words, a randomly taken image has nothing to do with any image of a dog, cat, boat, coffee machine, or whatever. An image of white noise, by definition, has no structure. So, functions that attempt to approximate $f(x) = y$ must do so well in a very small space of images. Be aware that even if they are small compared to the total dimension of the space, the islands Ω_y can still have large dimensions, meaning that you cannot decompose $f(x) = y$ into small, low-dimensional problems $f_k(x) = y_k$ with $x \in \Omega_{y_k}$.

So, the theorems that attempt to analyze (constrain) functions across the entire

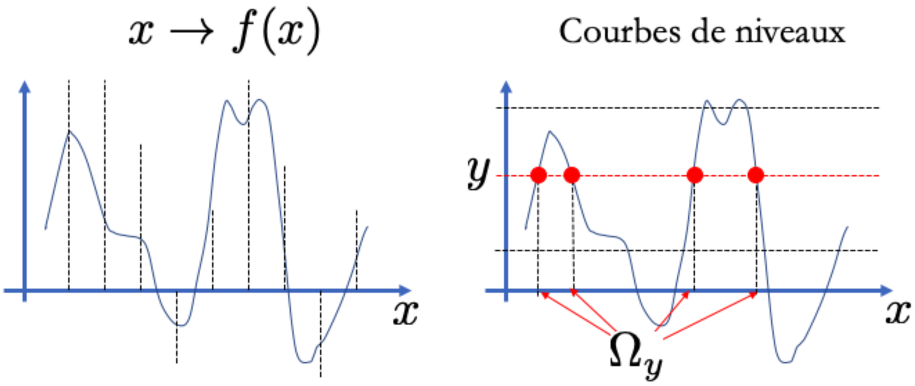

FIGURE 23 – The two perspectives on the problem $y = f(x)$: either we attack the problem from a viewpoint where for each value of x, we want to know $f(x)$ somewhat independently of whether x is a relevant signal (left diagram), or we study the level curves of f (right diagram) and focus on the values of x that ultimately matter.

space $[0,1]^d$ ask a question that doesn't really make sense because what needs to be analyzed is the restriction of f to the support of Ω_y. *First*, this is the data support, where learning can be attempted, and *second*, this is also the support of predictions where we want the approximation to be good.

The idea of this approach is not to look at functions that are more or less regular over the entire space $[0,1]^d$, but rather to focus on the *geometry* and the *location* of the islands Ω_y, which can be of large dimension. This is precisely **the perspective of algorithms and the Bayesian perspective**. Be aware that this is not a matter of *probabilistic versus deterministic* that differentiates the Bayesian perspective from the previous one. The idea is that instead of asking "*for any value of x, what is the value of the function $f(x)$?*", we will study the function through its *level sets $f = cte$* (Fig. 23). These different perspectives on a subject are similar to when you look at Riemann integration (the first perspective) and Lebesgue integration (the second perspective).

The Bayesian perspective is as follows: I want to know the label y given that I know x. So, the object is the conditional probability $p(y|x)$, and in the case of classification, for a given x_0, we take the y for which the probability $p(y|x_0)$ is maximum. This is the **maximum likelihood** (see the discussion in Course 2018 Sec. 7.2.1). Now, Bayes' theorem tells us that:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{110}$$

and we want to find the maximum with respect to y, which means the approximate value $\tilde{y}$ is obtained by:

$$\tilde{y} = \underset{y}{\operatorname{argmax}} \frac{p(x|y)p(y)}{p(x)} = \underset{y}{\operatorname{argmax}}\ p(x|y)p(y) \tag{111}$$

With $p(y)$ being the *prior*, which is prior information about class occurrence. In the following, let's assume that all classes are equiprobable (the volumes of Ω_y are identical), so we can drop $p(y)$ to find $\tilde{y}$. Finally, we can maximize the logarithm, which gives:

$$\tilde{y} = \underset{y}{\operatorname{argmax}}\ \log p(x|y) \tag{112}$$

So, in the Bayesian perspective, **we need to model not $f(x)$ but** $\log p(x|y)$ **or** $p(x|y)$, which means asking, given the label y, what is the location (geometric) of x in the space $[0,1]^d$? Which is precisely the location of Ω_y. Now, in high dimensions, there are **concentration phenomena** where the probability $p(x|y)$ is truly maximal on small domains Ω_y, and most images do not belong to any class due to a lack of structure. In fact, technically, what is more pertinent to study is the difference in probability between two classes:

$$\log p(x|y) - \log p(x|y') \tag{113}$$

Because if this difference is positive, we assign label y, otherwise label y' is the answer.

In doing so, the Bayesian approach is truly different because it's not f that we will focus on (e.g., regularity shape) and perform harmonic analysis, for example, as what's important is to ask where its support is and how to characterize it. In fact, through $p(x|y)$, it's **the modeling of x in each of the classes** that underlies. Thus, in high dimensions, the classification problem and the modeling problem of x are essentially the same, or more precisely, **we must tackle the problem of modeling the specificities** of x in Ω_y compared to the specificities of x in $\Omega_{y'}$. So, even when dealing with a **classification/regression problem** of the type $y = f(x)$ where initially we focused on f, **ultimately, we return to studying** $x(u)$. Hence, the deepening of signal analysis that will occupy us. We will show how **the nonlinear perspective will help us perform these modeling**. However, before that, let's review the results obtained so far.

5.2.4 Summary

What we have seen so far is that we can obtain:

- either *non-adaptive linear approximations*, where essentially we perform projections onto linear spaces for which the best basis is obtained through PCA, which, for stationary processes, gives the Fourier basis;
- or *adaptive non-linear approximations*, and for this to work, we must ensure that in the chosen basis, the ℓ^α norms are finite to guarantee the decrease of the approximation error.

Ultimately, both of these perspectives are effective in low dimensions, but as the dimension grows, we encounter **the curse of dimensionality**. This means that even Barron's theorem is ultimately not suitable for practical cases. We need to adopt another perspective that brings us back to the study of x. Because the way this barrier can be overcome is not so much that the function f is extraordinarily regular on $[0,1]^d$, but rather that **its support**, i.e., the places where it is interesting to have a good approximation, **is highly concentrated**. **So, the challenge is to characterize this support**, which is certainly restricted but not of low dimension. Therefore, we will return to the RAP triangle to **understand this notion of regularity**.

5.3 Information Theory. Wavelet Bases

During our study of a single-layer neural network, we realized the need to return to the analysis of the signal x, even in the case of classification/regression. We will adopt a non-linear perspective because the resulting **signal modeling** will be of much higher quality than Fourier analysis. In this context, we will try to understand what type of basis will allow us to do much better. In particular, we will address the question of **compression** through **Information Theory**, which precisely explains the **concentration phenomena** encountered in the previous section.

To approach this subject, we will start from the RAP triangle. Recall that in the linear framework, we began with the concept of *regularity* of a function (Sec. 3.3): we were able to characterize a function as *uniformly regular* through its translation-covariant derivatives, which are diagonalizable in the Fourier basis. This allowed us to discuss *sparsity*

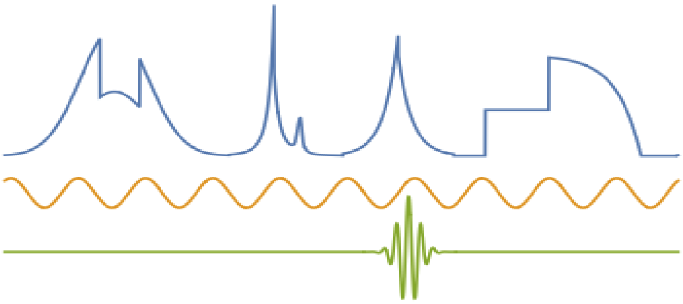

FIGURE 24 – Different perspectives on function analysis: either the linear framework, which assumes uniform regularity to give Fourier analysis with sinusoids delocalized in time/space, or the non-linear framework that studies non-uniformly regular functions with Wavelet analysis, which performs local analysis of transients/discontinuities.

and *low-dimensional approximation*. So, **the starting point in the linear context is the notion of uniform regularity**. However, as we have seen previously, signals that contain interesting *features* are those that exhibit discontinuities, transients, e.g., contours, changes in rhythms, the attack of a musical note, etc. S. Mallat tells us that if you change even just the first 50 milliseconds of a few seconds of a note produced by a violin in a musical piece to the beginning of the same note produced by a piano, then our perception is completely changed. This means that the perception of sound is strongly influenced by the signal's discontinuities, specifically the attack of a note by a violin or a piano. Therefore, we need to focus on a form of **piecewise regularity** that can represent a wide class of functions that are truly interesting for our concerns.

To do this, we need to find a way to localize the analysis of temporal transients in 1D (or spatial in 2D, etc), while in the Fourier framework, the cosines are localized in frequency but completely delocalized in time. Thus, we need localized sinusoids, namely **wavelets**[30] (Fig. 24).

5.3.1 Wavelet Analysis

A wavelet $\psi(u)$ is an oscillatory function with finite support, and since it is localized in space, it must be deformable not only by translation but also in scale to match the loca-

[30] Refer to the 2018 and 2020 courses.

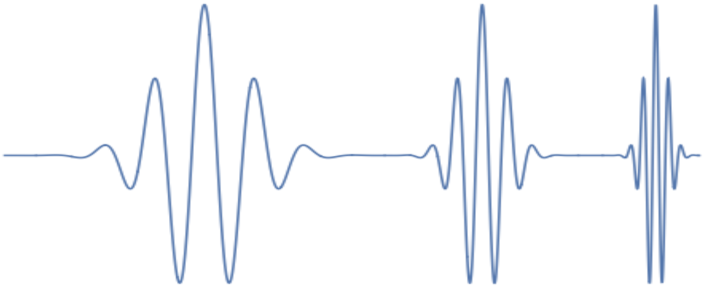

FIGURE 25 – Illustration of translation and scaling operations applied to a wavelet.

tion and size of the transient (Fig. 25). Thus, from $\psi(u)$, we introduce two transformations by defining the family $\{\psi_{v,s}\}$ (here in 1D):

$$\boxed{\psi_{v,s}(u) = \frac{1}{\sqrt{s}}\psi\left(\frac{u-v}{s}\right)} \tag{114}$$

Note the normalization $1/\sqrt{s}$ to obtain an orthonormal basis:

$$\|\psi_{v,s}\|^2 = \int |\psi_{v,s}(u)|^2 \; du = \|\psi\|^2 \tag{115}$$

We require that the wavelet oscillates, which is expressed by the constraint

$$\int \psi(u)du = 0 \tag{116}$$

Now, as we are interested in the analysis of the local regularity of the signal x using $\psi_{v,s}$, we define the **Wavelet Transform** as follows (wavelets are taken as real here):

$$\boxed{\begin{aligned} W_x(v,s) = \langle x, \psi_{v,s}\rangle = \int x(u)\psi_{v,s}(u) \; du = \int x(u)\frac{1}{\sqrt{s}}\psi\left(\frac{u-v}{s}\right) \; du \qquad (117) \\ = \int x(u)\tilde{\psi}_s(v-u) = (x * \tilde{\psi}_s)(v) \qquad (118) \end{aligned}}$$

with

$$\tilde{\psi}_s(u) = \frac{1}{\sqrt{s}}\psi\left(-\frac{u}{s}\right) \tag{119}$$

Thus, the Wavelet Transform can be viewed either as an operation of **inner product** between x and $\psi_{v,s}$ or as a **convolution** between x and the **filter** $\tilde{\psi}_s$ [31].

This operation measures **the local variation of** x **around** v **on a scale proportional to** s. So, concerning the RAP triangle, we will attempt to capture the regularity of x, specifically non-uniform regularity, and to do this, we will use sparsity to find a suitable basis for capturing local irregularities (even local discontinuities).

In fact, we don't have much choice when constructing the basis because we need a family of localized functions that can adapt to the scale/size of irregularities. Thus, almost mechanically, we arrive at the wavelet transform with the associated bases. Remember that, regarding the Fourier transform or the wavelet transform, these are not tools that you choose among others: the Fourier basis is "the" basis in the linear framework, with translation invariance as the underlying principle, just as for studying transient phenomena, we end up with wavelet bases.

However, to fully analyze the RAP triangle, we will need to define low-dimensional approximations. This type of approximation is easy to do as long as we have orthogonal bases. Therefore, the critical point is the construction of such bases from wavelets. Even before that, we need to demonstrate that we are effectively capturing the regularity of functions with these wavelets.

5.3.2 Local Lipschitz Regularity and Wavelet Coefficient Decay

Regarding local regularity, we will approach it in the sense of Lipschitz [32]:

Definition 1 *A function $x(u)$ is Lipschitz α at a point v if $\exists\, C > 0$ such that*

$$|x(u) - x(v)| < C|u - v|^{\alpha} \tag{120}$$

Figure 26 illustrates local variations of Lipschitz functions $0 \leq \alpha < 1$ (note that for $\alpha > 1$, the function is locally constant, and for a Brownian motion, we have $\alpha = 1/2 - \varepsilon$). If we

[31] Note that in the 2020 course, the perspective was that of convolutional filters, so the chosen normalization was $1/s$, while in the 2018 course, the factor was $1/\sqrt{s}$ because the "inner product" aspect was emphasized.

[32] Note: Rudolph O. S. Lipschitz (1832-1903), whose extension by Otto Ludwig Hölder (1859-1937) is used by S. Mallat.

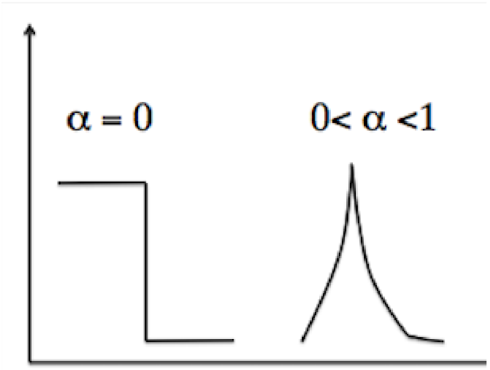

FIGURE 26 – Illustration of Lipschitz functions α for different values of α.

work on an interval, we can make the notion uniform as follows:

> **Definition 2** *A function $x(u)$ is uniformly Lipschitz α on an interval I if $\exists C > 0$ such that*
>
> $$\forall (u, v) \in I,\ |x(u) - x(v)| < C|u - v|^{\alpha} \tag{121}$$

What needs to be realized is that the value of α can be a constant in the uniform case, but more importantly, it can change locally and thus inform us about the **local regularity of the function**. Now, can we relate Lipschitz regularity to the Wavelet Transform?

In the case of Fourier, we were able to relate the regularity of a function to the rate of decay of inner products $|\langle x, e_n \rangle|$, so we looked at how the inner products evolve as the frequency ($\omega_n = 2\pi n$) increases. With wavelets, we will focus on a point v, and we will also increase the "frequency" of oscillation, where **the equivalent of ω_n is $1/s$, which is the scale change**. Indeed,

$$\widehat{\psi_{v,s}}(\omega) = \sqrt{s} e^{-i\omega v} \hat{\psi}(s\omega) \tag{122}$$

so if $|\hat{\psi}(\omega)|$ has a maximum at ω_0, $|\widehat{\psi_{v,s}}|$ has it at ω_0/s. Thus, by letting s tend to 0, we should be able to appreciate the rate of decay of inner products at high frequency and deduce the Lipschitz regularity α of the function. In fact, we have the following theorem:

Theorem 11 *If x is Lipschitz α at v, then $\exists C$ such that*

$$\forall s,\ |\langle x, \psi_{v,s}\rangle| < C s^{\alpha+1/2} \tag{123}$$

This theorem tells us that the more regular the function is, i.e., the closer α is to 1, the faster the inner product decay $|\langle x, \psi_{v,s}\rangle|$ as $s \to 0$, or equivalently at high frequencies. This is a different but related idea to what we had with Fourier. The main difference is that the **rate of decay is analyzed at a specific point (v)**, whereas in the Fourier case, the decay is analyzed over an entire interval, to the extent that if there is only one discontinuity in the interval, the decay is fixed at $1/\omega$, revealing nothing about the fact that the function may be 100 times differentiable outside of that discontinuity. **With the Wavelet Transform, we have a tool like a microscope searching for (ir)regularities locally**.

Is Theorem 11 an equivalence? In fact, there are several theorems. For example, *there is equivalence if, for instance, we have a function that is uniformly Lipschitz α over an arbitrarily small interval*. However, if we insist on the fact that we have a pointwise regularity, then the answer is that the equivalence is "almost true", but we need to change the bound by introducing a logarithmic correction because there can be issues with fractals. This is a result from 1990 in "micro-localization" (see S. Jaffard [33]). However, Theorem 11 with the sufficient condition is sufficient for our purposes.

Proof 11. To re-establish the result that pointwise Lipschitz condition imposes a constraint on the inner product, we can write it as

$$\langle x, \psi_{v,s}\rangle = \int x(u) \frac{1}{\sqrt{s}} \psi\left(\frac{u-v}{s}\right)\, du \tag{124}$$

and we know that the integral of the wavelet ψ is zero, which is also true for $\psi_{v,s}(u)$ as a

[33] For example, Theorem 3.1 in http://www.ens-lyon.fr/DI/wp-content/uploads/2009/07/Jaffard-IC2.pdf.

function of u, so

$$\begin{aligned}
|\langle x, \psi_{v,s}\rangle| &= \left|\int (x(u) - x(v)) \frac{1}{\sqrt{s}} \psi\left(\frac{u-v}{s}\right)\, du\right| \\
&\leq \int |x(u) - x(v)| \frac{1}{\sqrt{s}} \left|\psi\left(\frac{u-v}{s}\right)\right|\, du \\
&\leq C \int |u-v|^{\alpha} \frac{1}{\sqrt{s}} \left|\psi\left(\frac{u-v}{s}\right)\right|\, du \\
&\leq C \int |su'|^{\alpha} \sqrt{s} |\psi(u')| du' \\
&\leq C s^{\alpha+1/2} \int |u|^{\alpha} |\psi(u)| du
\end{aligned}$$

Now, since the wavelet ψ is localized, the right integral is a constant, so the inner product is indeed bounded, as indicated by the theorem. ■

Note that the proof relies on the fact that the wavelet is both *oscillatory* and *localized*, and the understanding of the theorem goes far beyond a simple change of variables: if the function is Lipschitz α, then the increments around v are multiplied by s^{α}, which is reflected in the constraint on the wavelet coefficients, i.e., the inner products.

6. Lecture 10 Feb.

At the end of the last session, we saw how the *local Lipschitz regularity* α of the signal can be captured in the coefficients of the wavelet transformation. Thus, we have a way to quantify the regularity of the signal x, and if it doesn't have too many discontinuities, then we can build *sparse representations* with *orthonormal wavelet bases.* Subsequently, from these representations, we will be able to develop *low-dimensional approximations*, completing the analysis of the RAP triangle that we embarked on in the *non-linear framework* to adapt to the signal x and understand the phenomena of concentration of level set supports $f(x) = y$ in high dimensions.

6.1 Lipschitz α Regularity and Scalogram

So, consider the signal $x(u)$. If we want to know the regularity around v_0, we can approximate it with the best polynomial approximation and study the approximation error.

> **Definition 3** *(**Lipschitz** α)*
> *Let $x(u)$ be such that*
>
> $$\exists C > 0 \text{ s.t. } \forall u \quad |x(u) - p_{v_0}(u)| \leq C|u - v_0|^{\alpha} \tag{125}$$
>
> *with $m - 1 \leq \alpha \leq m$ and p_{v_0} a polynomial of degree $m - 1$, then we say that x is a Lipschitz α function at v_0.*

Thus, **we can view the local regularity of a function as the error of a polynomial approximation**. This brings us back to the result of Taylor's expansion: if x is m times differentiable, then the Taylor remainder is $o((u - v_0)^m)$. Lipschitz regularity is an extension when α is a real number. Now, if $x(u)$ is a well-behaved function, α is relatively large, and the polynomial approximation is more effective, so we can approximate x with few parameters. But if α changes depending on v_0, how can we still obtain a sparse representation?

In the last session, we saw Theorem 11, which allows us to encode Lipschitz regularity α in the decay of wavelet coefficients. Figure 27 shows an example of a signal $x(u)$ with irregularities at several locations, and the result of the wavelet transform, where intensity in color represents $|\log(W_x(v, s))|$ with v on the x-axis and scale s on the y-axis, where $s = s_0\, 2^j\, 2^{n/Q}$ (octave j, voice n) with $Q = 16$ voices per octave, and s_0 as the smallest scale of the wavelet. This is known as a **scalogram**. High frequencies (small scales) are at the top, and low frequencies (large scales) are at the bottom. The effect of signal discontinuities can be clearly seen. The Morlet real wavelet is used[34].

Recall that the wavelet ψ has a zero mean (Eq. 116), so its Fourier spectrum is that of a **band-pass filter** (Fig. 28). Furthermore, we impose ψ to have m **zero moments**, i.e.,

[34] It is given by the function $\psi(x) = \pi^{-1/4} e^{-x^2/2} (\cos(\pi x (2/\log 2)^{1/2}) - e^{-\pi^2/\log 2})$, where the constant ensures zero integral, and its Fourier transform is $\hat{\psi}(\omega) = 2^{3/2} \pi^{1/4} e^{-\omega^2/2} e^{-\pi^2/\log 2} \sinh^2(\omega\pi/\sqrt{\log 4})$.

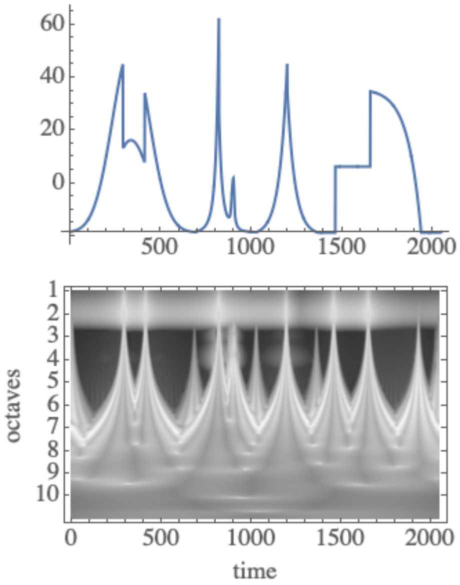

FIGURE 27 – Example of a Scalogram (bottom) obtained from the wavelet transformation of the signal in the top diagram. The x-axis represents the variable u (e.g., time), and the y-axis represents the scale $s = s_0\, 2^j\, 2^{n/Q}$ with j as the octave, n as the "voice", and $Q = 16$ as the number of voices per octave, with s_0 as the smallest wavelet scale. The intensity of the grays indicates the value of $|\log(W_x(v, s))|$ on a scale where 1 corresponds to black and 0 to white.

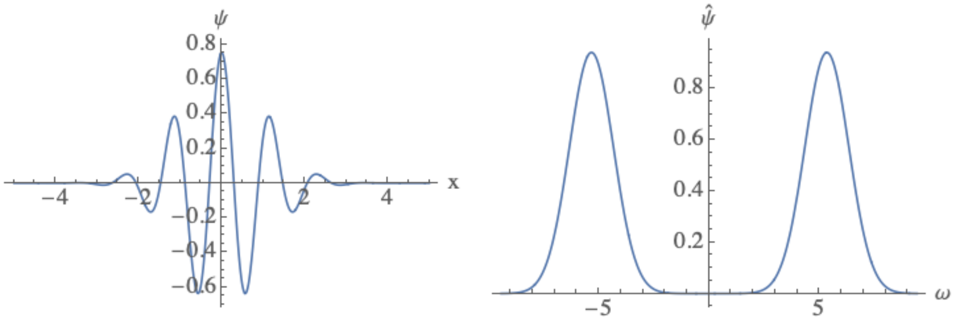

FIGURE 28 – The wavelet ψ has zero mean, making it a band-pass filter. Illustrated with the real Morlet wavelet.

$$\forall k,\ 0 \leq k < m, \quad \int \psi(u)u^k \, du = 0 \tag{126}$$

For instance, the Morlet real wavelet mentioned above has 1 extra zero moment. Therefore, if ψ has m zero moments, it naturally follows that for any polynomial $p(x)$ of degree $d < m$,

$$\int \psi(u)p(u)du = 0 \tag{127}$$

Why is it important to use such wavelets? The reason is that the wavelet "ignores" the polynomial part of $x(u)$ and is only sensitive to the polynomial approximation error. If the wavelet coefficient W_x is small, it signifies a small error, and vice versa. Therefore, in the scalogram shown in Figure 27, only the high-value coefficients that reflect coherent interference of the wavelet with the signal x in the vicinity of v and for a scale s ($\psi_{s,v}$) are visible. As the scale increases, the wavelet dilates, and it delocalizes the discontinuities, resulting in the appearance of "cones" that widen at the locations of v values where there are discontinuities. However, if the signal $x(u)$ has discontinuities almost everywhere, then the scalogram is filled with cones, as in Figure 29, and reading it becomes challenging, but **the figure reflects the regularity/irregularity of the signal $x(u)$ at all scales**.

So initially, we will analyze the properties of the scalogram, and then we will try to concentrate the maximum amount of information on a minimum number of coefficients, which will lead to a sampling of the scalogram resulting in orthonormal wavelet bases.

6.2 In-Depth Study of the Scalogram

Let's revisit Theorem 11 to provide a more precise version. This version was proven by S. Jaffard and characterizes the **pointwise regularity of the function**, shedding light on the famous **cones**. Here it is:

> **Theorem 12** ***(S. Jaffard)***
> *Let ψ be a wavelet with m zero moments. If x is Lipschitz $\alpha \leq m$ at a point v_0, then $\exists C > 0$ such that*
>
> $$|W_x(v,s)| \leq C s^{\alpha+1/2}\left(1 + \left|\frac{v-v_0}{s}\right|^{\alpha}\right) \tag{128}$$

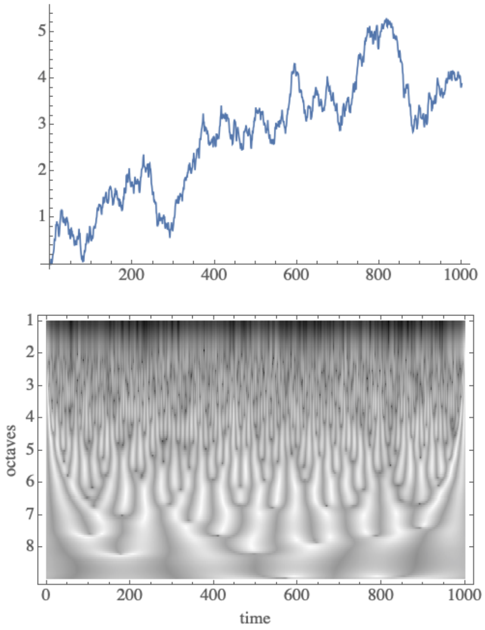

FIGURE 29 – Scalogram of a Brownian motion-like process.

> *and* ***conversely***[a]*, if* $\alpha' > \alpha$*, then we have*
>
> $$|W_x(v,s)| \leq C s^{\alpha+1/2}\left(1 + \left|\frac{v-v_0}{s}\right|^{\alpha'}\right) \tag{129}$$
>
> *then* x *is Lipschitz* α *at* v_0*.*
>
> [a] Note the difference in the exponent regarding the cone.

This result tells us that as s approaches 0 (i.e., high frequencies), the decay depends on α. In Figure 27, at a large scale, the function has smooth variations between its minimum and maximum, and the coefficients are fairly uniform along the u axis. However, as the scale decreases, the localization of discontinuities becomes more prominent.

The cone is defined by $|v - v_0|/s \leq 1$ because then the decay rate is dominated by $s^{\alpha+1/2}$, and the power α of $|v - v_0|/s$ controls cases of singularity like $\sin(1/u)$. Let's see how we can prove the first part of the theorem; the second, more technical part, will not be addressed here[35].

[35] See Proposition 3.2 of the document mentioned in the previous footnote about S. Jaffard, as well

Proof 12.
Let's express the coefficient $W_x(v,s)$ as follows:

$$W_x(v,s) = \int x(u)\frac{1}{\sqrt{s}}\psi\left(\frac{u-v}{s}\right)\,du$$

Now, the signal $x(u)$ is approximated by a polynomial $p_{v_0}(u)$ of degree $m-1$. Given that ψ has m zero moments, we know that $\int p_{v_0}(u)\psi(u)du = 0$. Note that this property also extends to $\psi_{v,s}(u)$ simply by a change of variable. So, with the same logic as the proof of Theorem 11, we can write

$$W_x(v,s) = \int (x(u) - p_{v_0}(u))\frac{1}{\sqrt{s}}\psi\left(\frac{u-v}{s}\right)\,du$$

which shows that the wavelet is not sensitive to the polynomial regularity of x. Thus,

$$\begin{aligned}|W_x(v,s)| &\leq \int |(x(u) - p_{v_0}(u))|\frac{1}{\sqrt{s}}\left|\psi\left(\frac{u-v}{s}\right)\right|\,du \\ &\leq C\int |u-v_0|^\alpha|\frac{1}{\sqrt{s}}\left|\psi\left(\frac{u-v}{s}\right)\right|\,du \\ &\leq Cs^{1/2}\int |su'+v-v_0|^\alpha|\psi(u')|\,du'\end{aligned}$$

Now, for any a and b, we can show that $|a+b|^\alpha \leq 2^\alpha(|a|^\alpha + |b|^\alpha)$, so

$$\begin{aligned}|W_x(v,s)| &\leq 2^\alpha Cs^{1/2}\int (|su'|^\alpha + |v-v_0|^\alpha)|\psi(u')|\,du' \\ &\leq 2^\alpha Cs^{1/2+\alpha}\left(\int |u'|^\alpha|\psi(u')|\,du' + \left|\frac{v-v_0}{s}\right|^\alpha\int |\psi(u')|\,du'\right)\end{aligned}$$

We have two integrals here that depend on the wavelet ψ, and these are constants that can be taken as the maximum and factored out, which concludes the proof. ■

To prove the "converse", one needs to start from the wavelet coefficients W_x and reconstruct the function x which is primarily done by showing that **the Wavelet Transform is invertible**. We will explore this fundamental point later (see also the 2020 course).

as links to the chapters in S. Mallat's book.

So, the theorem gives us a way to interpret the scalogram. However, it must be acknowledged that this image is far from being a sparse representation because we started with a 1D signal to obtain a 2D image. The question arises as to whether we can compress the information. Jean Morlet and Alex Grossmann played a pioneering role in the development of Continuous Wavelet Transform. J. Morlet (1931-2007), an engineer at Elf-Aquitaine, studied geological layers through the analysis of seismic waves, while A. Grossmann (1930-2019), a Franco-Croatian physicist, saw parallels with coherent states in quantum physics. This led a multitude of mathematicians from different fields to converge and study the Wavelet Transform, which analyzes a signal at different scales (see courses from previous years). In particular, how can we discretize the representation?

6.3 Towards a Sparse Representation: Dual Discretization

6.3.1 Scale Discretization

First, let's fix a discrete scale[36] with $s = 2^j$. We will show that, under a condition on the wavelet, this type of discretization is sufficient. In other words, having the coefficients $W_x(v, 2^j)$ allows us to recover x. Recall that $W_x(v, 2^j)$ can be seen as a convolution with the filter $\tilde{\psi}_{2^j}$ (Eq. 119). Now, when we talk about convolution, we bring Fourier Transform into the picture, so

$$\widehat{W_x}(\omega, 2^j) = \widehat{x}(\omega)\ \widehat{\tilde{\psi}_{2^j}}(\omega) \tag{130}$$

with

$$\widehat{\tilde{\psi}_{2^j}}(\omega) = \sqrt{2^j}\ \hat{\psi}^*(2^j\omega) \tag{131}$$

The question then becomes, can we reconstruct $\widehat{x}(\omega)$ from the coefficients $\widehat{W_x}(\omega, 2^j)$? This is only possible if the Fourier spectrum is fully covered by the filters $\hat{\psi}(2^j\omega)$. For the base wavelet ($j = 0$), the filter is that of a band-pass filter. Thus, it suffices for the supports of the dilated/contracted filters to gradually overlap so that if we take all $j \in \mathbb{Z}$, the entire Fourier spectrum is covered, i.e., without gaps. An illustration of how the filters $\hat{\psi}(2^j\omega)$

[36] This discretization is implicit in the images of the scalograms presented in Figures 27 and 29, as in practice, we only have a finite sample of $x(u)$ values.

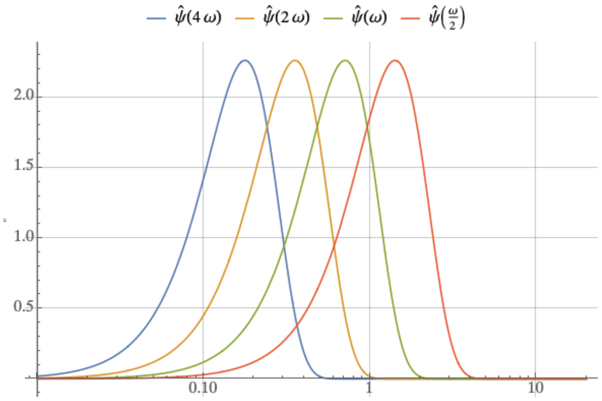

FIGURE 30 – Illustration of the evolution of the support of the band-pass filter given by $\hat{\psi}(2^j\omega)$: for $s = 2^j > 1$, the support shifts towards lower frequencies, and for $s < 1$, it shifts towards higher frequencies.

evolve for different values of j [37] is shown in Figure 30.

The following theorem allows us to formalize this concept [38]:

Theorem 13 *If we have the following Littlewood-Paley condition*

$$\forall \omega, \sum_{j\in\mathbb{Z}} |\hat{\psi}(2^j\omega)|^2 = 1 \tag{132}$$

then

$$x(u) = \sum_{j\in\mathbb{Z}} 2^{-j}(\widehat{W_x}(\omega, 2^j) * \psi_{2^j})(u) \tag{133}$$

The proof is straightforward as usual, using the Fourier transform of the convolution product. Thus, **provided that we cover the Fourier spectrum of the signal well, we can restrict ourselves to using only dyadic scales** $s = 2^j$ $(j \in \mathbb{Z})$. But we want to go further, namely, **discretize the axis u, meaning perform a sampling** of the signal.

[37] This is the wavelet $\psi_\sigma(u) = 2\pi^{-1/4}/\sqrt{3\sigma}(1 - t^2/\sigma^2)e^{-1/2(t/\sigma)^2}$ whose Fourier transform is given by $\hat{\psi}_\sigma(\omega) = 2\sqrt{2/3}\pi^{1/4}\sigma^{5/2}\omega^2 e^{-1/2(\sigma\omega)^2}$. For the illustration, $\sigma = 2$.

[38] Refer to Courses 2018 and 2020.

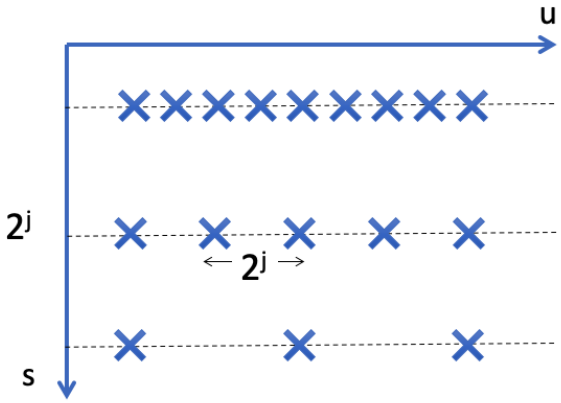

FIGURE 31 – Optimal sampling along the "space/time" axis in accordance with dyadic scale discretization.

6.3.2 Discretization of the "Space" Variable

When we write $W_x(v, 2^j)$, the variable v, which is akin to space or time, is continuous. How can we discretize it? Let's use the signal x's filtering formulation again, namely,

$$W_x(v, 2^j) = (x * \tilde{\psi}_s)(v) \tag{134}$$

And the question is how to judiciously choose the values of v to lose no information? Do we have an intuition on this? The answer is yes because the size of the wavelet along v is proportional to 2^j, so we need to use a sampling that's also proportional to 2^j to cover the entire support of x properly. In fact, the result is more precise because we will show that the sampling interval is equal to 2^j, and hence the samples are $v_n = 2^j n$ (Fig. 31). So, we will have the coefficients $W_x(2^j n, 2^j)$ obtained using wavelets $\psi_{2^j n, 2^j}$, but to simplify notation, we will write $\psi_{j,n}$. If we use the "inner product" perspective, we know that

$$W_x(2^j n, 2^j) = \langle x, \psi_{j,n} \rangle \tag{135}$$

and the question is whether the family of functions $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$ is **an orthonormal basis**, for example, of $L^2(\mathbb{R})$ or $L^2([0,1])$), depending on the problem? Because then we can reconstruct the signal.

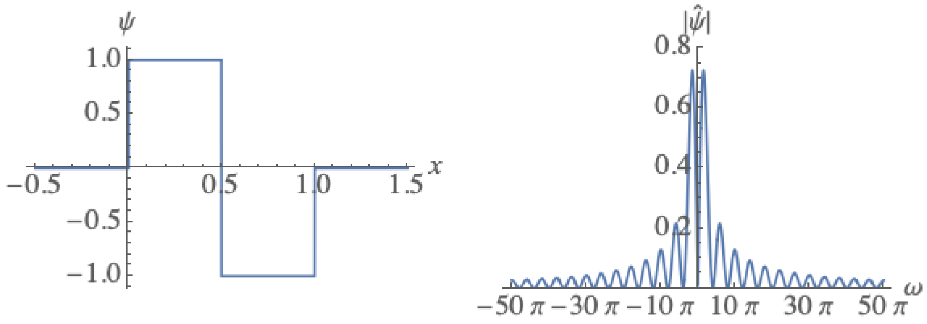

Figure 32 – Haar wavelet (also "Db1" in the Daubechies wavelet family) constructed in real space. Decays as $1/\omega$ in Fourier space.

6.3.3 Orthonormal Bases?

The question of finding orthonormal bases of this kind is not new, as **Alfred Haar** (1885-1933), a Hungarian mathematician, gave an example in 1909 (Fig. 32). It's quite easy to see that the inner product of two Haar wavelets is zero, and thus, it naturally follows that the family $\{\psi_{j,n}^{Haar}\}_{(j,n)\in\mathbb{Z}^2}$ forms an orthonormal basis of $L^2(\mathbb{R})$.

In 1948-49, **Claude Shannon** (1916-2001) and **Harry Nyquist** (1889-1976) proved a well-known theorem[39] known as the *Shannon Sampling Theorem*. Even though Shannon didn't mention any wavelets, he used a **perfect band-pass filter** (Fig. 33). It is evident that the filters $\hat{\psi}_j^{Sha}(\omega) \propto \hat{\psi}^{Sha}(2^j\omega)$ have supports $[2^{-j}\pi, 2^{-j+1}\pi]$ (similarly for negative frequencies). It is, therefore, easy to obtain the Littlewood-Paley condition (Th. 13). Thus, we can reconstruct the signal, which is also demonstrated in another way by the sampling theorem.

So, these two examples were known for a very long time, and the question that remained unanswered for a long time was whether there are other types of functions that have these properties. However, the two examples have disadvantages: Haar's wavelet is discontinuous and has no zero moments beyond its zero integral, so it cannot capture polynomial regularities. On the other hand, Shannon's wavelet in real space[40] decays as

[39] The *Whittaker–Nyquist–Kotelnikov–Shannon* theorem, to be more comprehensive about the history of the theorem.

[40] According to the definition of the Fourier transform, we find $\psi^{Sha}(u) = \operatorname{sinc}(t/2) - 2\operatorname{sinc}(t)$ with

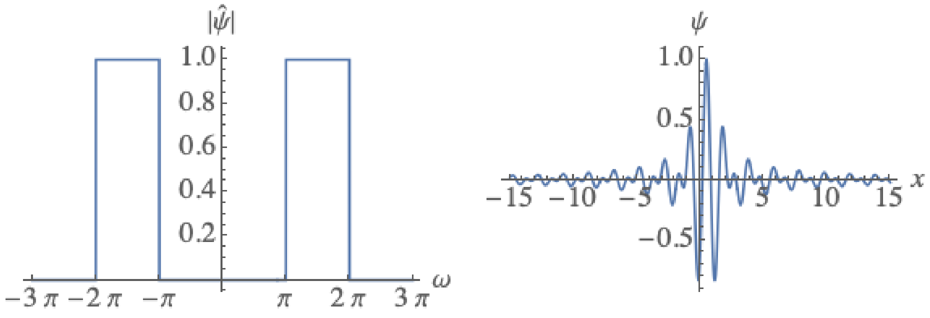

FIGURE 33 – Shannon wavelet constructed in Fourier space. Decays as $1/u$ in real space.

$1/u$, which is not a good indication of spatial localization. In fact, we would like to have wavelets that are well localized in space, highly regular with a sufficient number of zero moments, and, as a bonus, localized in Fourier space as well. Note that you cannot localize equally well in both spaces due to the **Uncertainty Principle** (See the 2020 course).

It must be said that it seemed impossible to satisfy all these constraints. In fact, Roger Balian and Francis Low, two theoretical physicists, had shown that it was not possible to construct orthonormal bases of $L^2(\mathbb{R})$ in the form of $g_{m,n}(u) = e^{2\pi imu}g(u-n)$ with $(m,n) \in \mathbb{Z}$ that are both localized in real space and Fourier space. It was a total surprise, and the remarkable result of **Yves Meyer** in 1986 was to find such a family of functions while trying to prove that it wasn't possible!

Yves Meyer's wavelet is both C^∞ and rapidly decreasing. It is constructed in Fourier space as follows:

$$\hat{\psi}^{Meyer}(\omega) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{i\omega/2} \sin\left[\frac{\pi}{2}\nu\left(\frac{3}{2\pi}|\omega| - 1\right)\right] & \frac{2\pi}{3} \leq |\omega| \leq \frac{4\pi}{3} \\ \frac{1}{\sqrt{2\pi}} e^{i\omega/2} \cos\left[\frac{\pi}{2}\nu\left(\frac{3}{4\pi}|\omega| - 1\right)\right] & \frac{4\pi}{3} \leq |\omega| \leq \frac{3\pi}{3} \\ 0 & \text{elsewhere} \end{cases} \tag{136}$$

$t = (2u-1)\pi$.

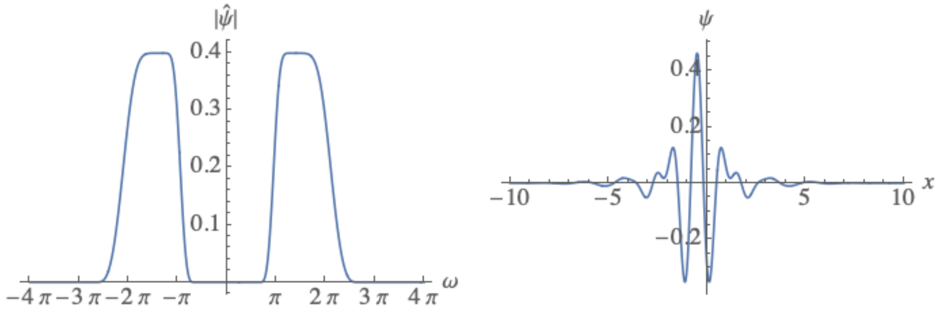

Figure 34 – Example of a Meyer wavelet constructed in Fourier space. Note its rapid decay in real space while being localized in Fourier space.

with $\nu(x)$ a C^k or C^∞ function satisfying

$$\nu(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x \geq 1 \end{cases} \qquad \text{and} \qquad \nu(x) + \nu(1-x) = 1 \tag{137}$$

An example is shown in Figure 34.

Following this result, a mathematical conceptual framework was developed that allowed all these results to come together and construct new orthonormal bases that satisfy various criteria of regularity and localization in both real and Fourier spaces. This is the **multiresolution analysis**[41].

However, to develop this framework of **orthonormal wavelet bases**, we need to introduce results from signal processing. Firstly, the **Shannon Sampling Theorem**, but in a more general form than what we are used to. Its generalization will lead us to **multiresolutions**. The general idea is as follows: **wavelets provide details of the analyzed function**, and therefore, we can improve the approximation of this function by gradually aggregating finer and finer details. Multiresolution analysis naturally leads to the realm of **multigrid approximations**, which are also found in numerical analysis and probability. This will lead to the construction of orthonormal bases, and we will link this to **filter bank algorithms**, which have been found to be **the basis of deep neural networks**, except that

[41] Refer to the 2018 course for an introduction to MRA or AMR, and the 2020 course for another aspect.

these networks include a fundamental non-linearity (see the 2020 course). Finally, we will arrive at low-dimensional approximations and the notion of sparsity, which was the initial motivation. Remember: if a function is regular, then you apply Fourier harmonic analysis; if the function has local discontinuities, you need to turn to multiresolution analyses.

6.4 Shannon Sampling Theorem

First, let's establish the following result, which states that **sampling a signal is equivalent, in the Fourier domain, to periodicization**:

> **Theorem 14** *If $\{x(nT)\}_n$ is a sample of the signal $x(u)$ ($u \in \mathbb{R}$), then the Fourier series*
> $$\sum_{n\in\mathbb{Z}} x(nT)e^{-inT\omega} = \frac{1}{T}\sum_{k\in\mathbb{Z}} \hat{x}\left(\omega - \frac{2k\pi}{T}\right) \tag{138}$$

This is a fundamental theorem in signal processing. Its proof can be seen as a consequence of the *Poisson summation formula*[42].

Proof 14. Let $\hat{a}(\omega)$ be the function on the right-hand side of the equality, which is $2\pi/T$-periodic. Therefore, we can write it as

$$\hat{a}(\omega) = \sum_{n\in\mathbb{Z}} a(n)e^{-inT\omega}$$

[42] We can formulate the equality, based on the definition of the Fourier transform used in this document, as $\sum_{n\in\mathbb{Z}} f(nT) = 1/T\sum_{k\in\mathbb{Z}} \hat{f}(2\pi k/T)$, which relates sampling in the real space and the Fourier space. Additionally, we can observe that $e^{-inT\omega} = \int_{-\infty}^{\infty} \delta(u-nT)du$, allowing us to work with the equality of two Fourier transforms by considering distributions, especially $T\sum_n \delta(u-nT) = \sum_k e^{i2k\pi u/T}$.

and we need to demonstrate that the $a(n)$ are the same as the $x(nT)$ on the left-hand side of the equality. The $a(n)$ are given by

$$\begin{aligned}
a(n) &= \frac{T}{2\pi}\int_0^{2\pi/T} \hat{a}(\omega)e^{inT\omega}\,d\omega \\
&= \frac{T}{2\pi}\int_0^{2\pi/T} \frac{1}{T}\sum_{k\in\mathbb{Z}} \hat{x}\left(\omega - \frac{2k\pi}{T}\right) e^{inT\omega}\,d\omega \\
&= \frac{1}{2\pi}\sum_{k\in\mathbb{Z}}\int_0^{2\pi/T} \hat{x}\left(\omega - \frac{2k\pi}{T}\right) e^{inT\omega}\,d\omega \\
&= \frac{1}{2\pi}\sum_{k\in\mathbb{Z}}\int_{2k\pi/T}^{(2k+1)\pi/T} \hat{x}(\omega')e^{inT\omega'}\,d\omega' \\
&= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \hat{x}(\omega')e^{inT\omega'}\,d\omega' = x(nT)
\end{aligned}$$

(Note: The inversion of sum and integral can be performed if the sum $\sum_{k\in\mathbb{Z}} \hat{x}\left(\omega - 2k\pi/T\right)$ converges uniformly, which requires some regularity assumptions on $x(u)$ that we assume to be satisfied.) ■

The sampling theorem allows us to reconstruct $x(u)$ from the $x(nT)$. However, it's easy to understand that if the function is highly irregular between the samples, we can't reliably reconstruct the signal. This necessitates **regularity assumptions**, which in turn imply **constraints on the decay of Fourier coefficients**. Let's imagine that the supports of the functions $\hat{x}\left(\omega - \frac{2k\pi}{T}\right)$ **do not overlap** (***no aliasing***), meaning that the support of $\hat{x}(\omega)$ is contained within $[-\pi/T, +\pi/T]$, which is a strong constraint. In that case, we can perform **low-pass filtering**, which in the spatial/temporal domain is equivalent to **convolution** with a ***sinc cardinal*** function.

Theorem 15 *(**Shannon**)*
If the support of $\hat{x}(\omega)$ is included in $[-\pi/T, +\pi/T]$, then

$$x(u) = \sum_{n\in\mathbb{Z}} x(nT)\,\phi_T(u - nT) \tag{139}$$

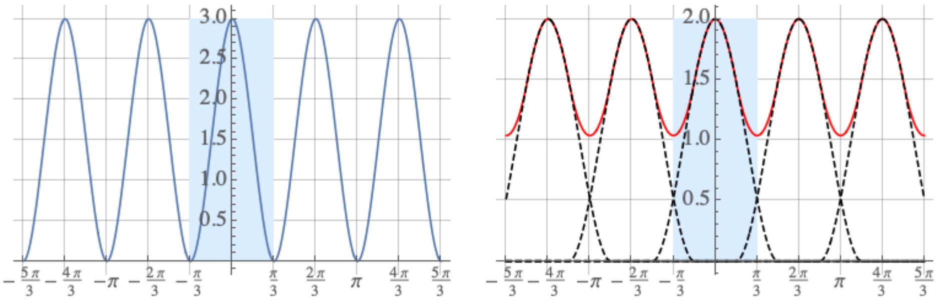

FIGURE 35 – Fourier spectra with replications every $2k\pi/T$, where on the left there is no aliasing, unlike the case shown on the right (here $T = 3$). The baseband is illustrated by the blue rectangle.

with

$$\phi_T(u) = \frac{\sin(\pi u/T)}{\pi u/T} = \mathrm{sinc}(\pi u/T) \tag{140}$$

whose Fourier Transform is the ideal low-pass filter

$$\widehat{\phi_T}(\omega) = T\ \mathbf{1}_{[-\pi/T,+\pi/T]}(\omega) \tag{141}$$

NDJE: A proof proposition can be developed as follows, relying on Theorem 14. Let's take the Fourier Transform of the right-hand side, we have

$$\sum_{n\in\mathbb{Z}} x(nT)e^{-i\omega nT}\widehat{\phi_T}(\omega) = \sum_{k\in\mathbb{Z}} \hat{x}\left(\omega - \frac{2k\pi}{T}\right)\mathbf{1}_{[-\pi/T,+\pi/T]}(\omega)$$

Now, as the support of $\hat{x}(\omega)$ is included in $[-\pi/T, +\pi/T]$, it is also the case for $\omega - 2k\pi/T$, which mechanically constrains k to be 0. Thus, we recover $\hat{x}(\omega)$, the Fourier Transform of the left-hand side.

Let's view this classic theorem from a different perspective. It tells us that if the support is contained in $[-\pi/T, +\pi/T]$, everything works fine. But what happens when we're not in this case? As illustrated in Figure 35 (right), once we cut off the baseline $\omega \in [-\pi/T, \pi/T]$, the Fourier spectrum no longer matches that of $x(u)$, and we can't reconstruct the signal.

In signal processing, to avoid aliasing, we start by **pre-filtering the signal** $x(u)$ within

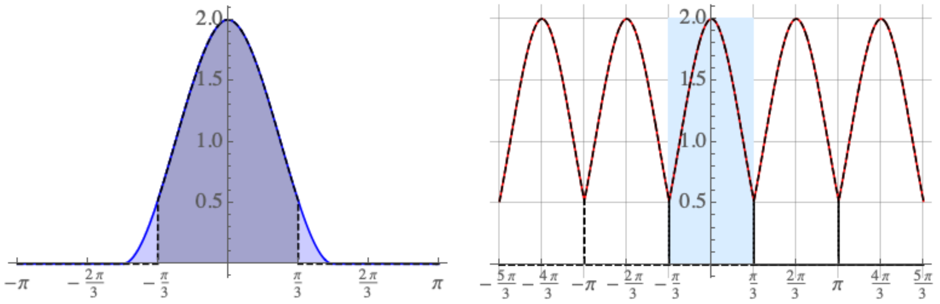

FIGURE 36 – To avoid aliasing, we pre-filter the signal within the $[-\pi/T, +\pi/T]$ band.

the $[-\pi/T, +\pi/T]$ band using the ideal low-pass filter $\phi_T(u)$:

$$x \xrightarrow{\text{pre-filtering}} (x * \phi_T)(u) = x_T(u) \tag{142}$$

This results in a signal $x_T(u)$ that no longer has high-frequency components. More precisely:

$$\widehat{x_T}(\omega) = T\ \hat{x}(\omega)\ \mathbf{1}_{[-\pi/T,+\pi/T]}(\omega) \tag{143}$$

meaning that the support of $\widehat{x_T}$ is well-contained within the interval $[-\pi/T, +\pi/T]$. Thus, we effectively avoid aliasing, as illustrated in Figure 36.

This low-pass pre-filtering process is an **example of linear approximation**. Essentially, we approximate the signal by an element of a linear space defined as

$$V_T = \{x \ / \ \text{Support } \hat{x} \subset [-\pi/T, +\pi/T]\} \tag{144}$$

This involves **interpolation** using the functions $\phi_T(u)$. In Figure 37, on the left side, you can see the blue signal $x(u)$, a sampling with $T = 1/2$, and the contribution $x(nT)\phi_T(u - nT)$ for $nT = 7$. Note that the zeros of $\phi_T(u - nT)$ are $u_k = kT$ with $k \in \mathbb{Z}^*$, so when we sum all the contributions, we obtain a function (red) that passes through all the points (red dots) of the sampling $x(nT)$. The result for $T = 1/2$ is shown as the green dashed curve on the right side of Figure 37. To illustrate the effect of the sampling interval, the approximation with $T = 1$ is also shown as the orange dashed line.

In fact, the family of functions $\{\phi_{T,n}(u) = \phi_T(u - nT)\}_{n\in\mathbb{Z}}$ forms an **orthogonal**

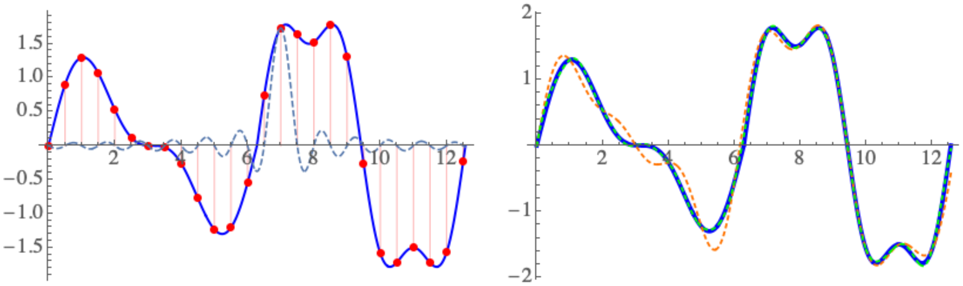

FIGURE 37 – Illustration of the interpolation formula Eq. 139 from Shannon's theorem (Th. 15). In blue, the signal $x(u)$ sampled with $T = 1/2$ (red points). The contribution $x(nT)\phi_T(u - nT)$ for $nT = 7$ is shown on the left, while the approximation resulting from the sum of all contributions is given on the right as green dashed lines (hardly distinguishable from the blue curve). The approximation of lower quality obtained with $T = 1$ is also shown as orange dashed lines.

basis of V_T. First, all functions are members of V_T because their Fourier support is that of $\widehat{\phi_T}$, thus contained within $[-\pi/T, +\pi/T]$. Second, the inner product between $\phi_{T,n}$ and $\phi_{T,m}$, as seen by Plancherel, is the integral of the product of their Fourier transforms, i.e.,

$$\langle \phi_{T,n}, \phi_{T,m} \rangle = \int \widehat{\phi_{T,n}}(\omega)\widehat{\phi_{T,m}}^*(\omega)\ d\omega \propto \int_{-\pi/T}^{\pi/T} e^{-i\omega T(n-m)}\ d\omega \propto \delta(n-m)$$

With this basis of V_T, we know that **the best linear approximation is the orthogonal projection** of x onto this space (Sec. 3.1), which naturally gives us the expansion:

$$P_{V_T}x(u) = \sum_{n\in\mathbb{Z}} \frac{1}{T} \langle x, \phi_{T,n} \rangle\ \phi_T(u - nT) \tag{145}$$

If $x \in V_T$, then x is equal to its orthogonal projection, which is the essence of Shannon's sampling theorem. Note that if we compare the two expansions on the basis of $\{\phi_{T,n}\}$ (Eq. 139), we realize that [43]

$$\langle x, \phi_{T,n} \rangle = T\ x(nT) \tag{146}$$

[43] where $T = \int \phi_T^2(u)du$.

Thus, we have highlighted another way to express **Shannon's sampling theorem, which is a very specific case of linear approximation** where the filter is that of an ideal low-pass filter. What we are going to see is that we can **generalize this result by using different types of filters**. The reason we want to change the filter is that Shannon's filter is discontinuous in Fourier, so it only decays as $1/u$ in the real space, which doesn't allow for good localization of the irregularities in the sought-after signal $x(u)$.

So, we will generalize the space V_T in such a way that we will define nested spaces. The analogy can be seen clearly in an image, where capturing essential structures is done by using versions of the initial image taken at different resolutions. Therefore, we need to set up a mathematical framework to understand these multi-scale structures. In this framework, Wavelets will naturally find their place, as the wavelet coefficients are the famous details of the image or solutions of a differential equation (ED). Moreover, if the image is piecewise regular, then **the representation is sparse**, as the largest coefficients are localized at the edges. In this scheme, we will also see that **cascade filter algorithms** naturally arise. These are at the core of deep neural networks.

7. Lecture 17 Feb.

During this session, we will build the foundations of Wavelets from Multiresolution Analysis. This topic, dear to S. Mallat, has many applications in mathematics and will be used to model the data $x(u)$ in order to understand the concentration phenomena in problems of the form $f(x) = y$, which are at the core of neural networks. So, for the record, the Wavelet Transform allows us to analyze the regularity of a function in a "space-scale" or "time-scale" plane through what is called the *scalogram*. Wavelet coefficients allow us to characterize the *local regularity* of the signal x at any point v. Large coefficients are found near singularities.

However, we have seen that we can compress the representation to make it sparse by first sampling the scale space ($s = 2^j$) to retain only the coefficients $W(u, 2^j)$ because if we have a wavelet whose dilated/contracted versions cover the Fourier space well (Littlewood-Paley condition of Theorem 13), then we can recover the signal. We also found by an intuitive argument that we can probably perform real-space sampling ($v_n = n2^j$). That

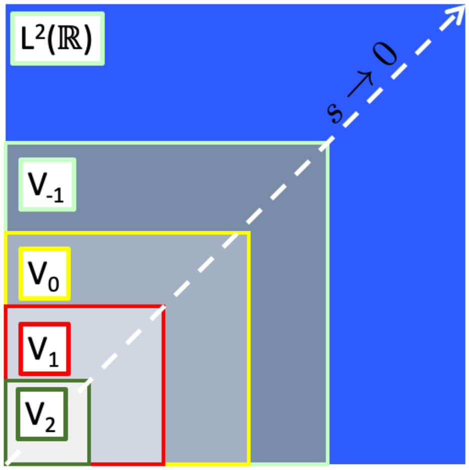

FIGURE 38 – Family of nested linear spaces V_i.

is, we wonder if the family of functions

$$\left\{ \psi_{j,n}(u) = \frac{1}{\sqrt{2^j}} \psi \left(\frac{u - 2^j n}{2^j} \right) \right\}_{(j,n) \in \mathbb{Z}^2} \tag{147}$$

is an **orthonormal basis** of the space L^2 of signals $x(u)$. If such a base(s) exists, then we can apply the previously used scheme to eliminate small wavelet coefficients to obtain a *nonlinear low-dimensional approximation*.

So, the problem we are addressing is to construct such orthonormal bases using multiresolutions. The ideas came in particular from the field of computer vision, in part due to limited memory size in the 1980s. Indeed, the question was whether it was possible to minimize the size of an image without losing essential information contained in the original image and to proceed with a progressive aggregation of details to refine the perception of the object under study. So, first, we will define the projection of the signal on grids at multiple resolutions, and then, by complement, we will add the details. In doing so, we will discover orthonormal bases along the way.

7.1 Multiresolutions

7.1.1 The Definition

We need to define projections at different scales, which are nested linear spaces (Fig. 38) [44].

Definition 4 *(**Multiresolution**)*
Let the family of linear spaces $\{V_j\}_{j\in\mathbb{Z}}$ indexed by j associated with scale 2^j be called a multiresolution if it satisfies the following properties:

i) We lose resolution when going from scale 2^j to 2^{j+1}, i.e.,

$$V_{j+1} \subset V_j \tag{148}$$

ii) We relate V_j and V_{j+1} according to the equivalence

$$x(u) \in V_j \Leftrightarrow x(u/2) \in V_{j+1} \tag{149}$$

we dilate $x(u)$ in one direction and contract it in the other.

iii) Moreover, if we approximate $x(u)$ by its orthogonal projection onto V_j, we would like that when the resolution is infinite ($s \to 0$, or $j \to -\infty$), we can completely reconstruct the signal, i.e.,

$$\lim_{j\to-\infty} \|x - P_{V_j}x\| = 0, \quad \forall x \in L^2(\mathbb{R}) \tag{150}$$

iv) Conversely [a], if $j \to +\infty$ or the scale becomes infinite, then we lose resolution and cannot reconstruct the signal

$$\lim_{j\to+\infty} P_{V_j}x = 0 \tag{151}$$

[44] For readers of the 2018 notes: for Section 6.6, one must be careful about how the indices j are arranged because they are opposite to the definition used here. This is a small gymnastics that must also be paid attention to in publications.

> *v) Finally, we need to add one last property related to the fact that we are building approximations on grids, and when we translate $x \in V_j$, we do not change the resolution, there is a translation invariance. Therefore,*
>
> $$\exists \phi \in V_0 \ / \ \{\phi(x-n)\}_{n\in\mathbb{Z}} \text{ is an orthonormal basis of } V_0 \tag{152}$$
>
> [a] These two properties can be written as: $\cup_{j\in\mathbb{Z}} V_j = L^2(\mathbb{R})$ and $\cap_{j\in\mathbb{Z}} V_j = 0$.

Property (v) can be extended to all spaces V_j by the following lemma:

> **Lemma 1** $\forall j \in \mathbb{Z}$, *then*
>
> $$\left\{\phi_{j,n}(u) = \frac{1}{\sqrt{2^j}}\phi\left(\frac{u-2^jn}{2^j}\right)\right\}_{n\in\mathbb{Z}} \tag{153}$$
>
> *is an orthonormal basis of V_j.*

7.1.2 Some Examples of Multiresolutions

To define the function ϕ, which is a cornerstone of the multiresolution definition, let's go back to the sampling theorem. Therefore, let

$$\phi(u) = \mathbf{1}_{[0,1[}(u) \tag{154}$$

We will see that ϕ is related to the Haar wavelet. It is quite clear that the family $\{\phi(u-n)\}_{n\in\mathbb{Z}}$ is an orthonormal basis of the linear space V_0 of piecewise constant functions with a step of 1. Similarly, $\phi_j(u) = \phi(u/2^j)$ is the function $\mathbf{1}_{[0,2^j[}(u)$, and the associated family is an orthonormal basis of the space V_j of piecewise constant functions with a step of 2^j, i.e., constant on the intervals $[2^jn, 2^j(n+1)[$. In this context, constructing an approximation of x at scale 2^j means obtaining the approximation by a step function over intervals of size 2^j. An example is shown in Figure 39 (left).

Another example is provided by defining $\phi(u)$ through its Fourier spectrum, as in Shannon's theorem 15:

$$\hat{\phi}(\omega) = \mathbf{1}_{[-\pi,\pi]}(\omega) \tag{155}$$

In real space, we find that $\phi(u) = \mathrm{sinc}(\pi u)$, and the decomposition of $x(u)$ corresponds to

$$x(u) = \sum_n \alpha_n \phi(u-n) \tag{156}$$

We know that $\alpha_n = x(n)$ according to Shannon, and now we can vary the scale parameter to obtain approximations at different resolutions. In the Fourier domain, $\hat{\phi}_j(\omega) = \mathbf{1}_{[-2^{-j}\pi, 2^{-j}\pi]}(\omega)$, so when $j \to -\infty$, we cover higher and higher frequencies, and we obtain higher-quality approximations. An example is shown in Figure 39 (right).

The two previous examples are extremes: in the first one, the function ϕ is discontinuous in space, while in the second one, it's the Fourier spectrum that is discontinuous. However, we would like to have more regularity in both spaces. Already at first glance, we can say that the piecewise constant approximation is very rough, and we could opt for at least a piecewise linear approximation, or even a piecewise polynomial one. Therefore, we think that we can certainly obtain higher-quality multiresolutions. What we will see is that with each **multiresolution**, we obtain a **wavelet basis**, the construction of which is based solely on that of **discrete filters**, and in particular (or above all) the one associated with the function ϕ. Finally, we will find **filtering-subsampling algorithms** that also appear in neural networks.

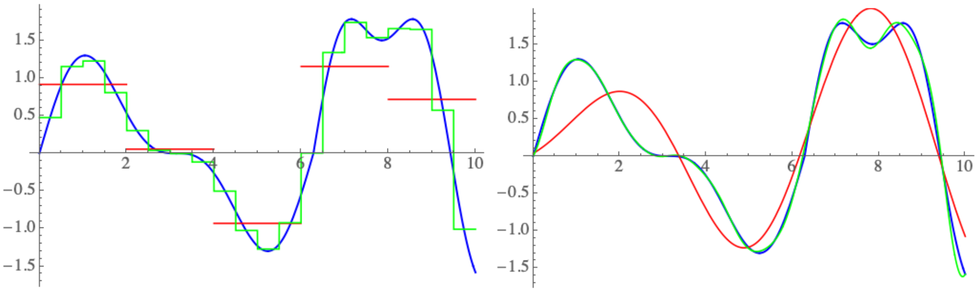

FIGURE 39 – Examples of multiresolution analysis with $\phi(u) = \mathbf{1}_{[0,1[}(u)$ (left) and $\phi(u) = \mathrm{sinc}(\pi u)$ (right): in red $j = 1$, in green $j = -1$, and in blue the function $x(u)$.

7.2 Filter Banks

The construction of filters, and more precisely, *filter banks*, which are networks of bandpass filters, is a traditional topic in signal processing aimed at separating the input signal into multiple components. In fact, wavelet theory has somewhat solidified empirical knowledge regarding the construction of filter banks.

First, let's ask the question: how do we construct an approximation P_{V_j}? It's an orthogonal projection onto the basis of V_j, so we have

$$P_{V_j}x = \sum_{n\in\mathbb{Z}} \langle x, \phi_{j,n}\rangle \phi_{j,n} \tag{157}$$

Now, the inner product of x with $\phi_{j,n}$ can be seen as a convolution. Let's introduce the function

$$\tilde{\phi}_j(u) = \frac{1}{\sqrt{2^j}}\phi\left(-\frac{u}{2^j}\right) = \phi_j(-u) \tag{158}$$

Thus, we can interpret the projection onto V_j as first a low-pass filtering with $\tilde{\phi}_j$ followed by interpolation with $\phi_{j,n}$:

$$\boxed{P_{V_j}x = \sum_{n\in\mathbb{Z}} \overbrace{\underbrace{(x * \tilde{\phi}_j)(2^j n)}_{\text{filtering}}\ \phi_{j,n}}^{\text{interpolation}}} \tag{159}$$

Finally, we can immediately generalize Shannon's theorem [45]:

> **Theorem 16** *(multiresolution sampling)*
> *If $x \in V_j$ of a multiresolution, then the orthogonal projection $P_{V_j}x = x$ (and vice versa), and thus*
>
> $$x(u) = \sum_{n\in\mathbb{Z}} (x * \tilde{\phi}_j)(2^j n)\ \phi_j(u - 2^j n) \tag{160}$$

Shannon's theorem becomes a special case when $\hat{\phi}(\omega)$ is the ideal low-pass filter on $[-\pi, \pi]$. However, what interests us is the transition between two different resolution grids, and

[45] Note that here, it's the same function ϕ that is used for filtering and interpolation. This scheme can be generalized by using two distinct functions.

this is where wavelets come into play. In doing so, we will focus on the two properties (i) and (ii) of a multiresolution to derive wavelets.

Let's fix $j = 0$. According to property (i), we know that $V_1 \subset V_0$ and $\phi(u) \in V_0$, so according to (ii), $\phi(u/2) \in V_1$, and therefore $\phi(u/2) \in V_0$. Now, since the $\{\phi(u-n)\}_n$ form a basis for V_0, we can decompose $\phi(u/2)$ using this basis[46]:

$$\boxed{\frac{1}{\sqrt{2}}\phi\left(\frac{u}{2}\right) = \sum_{n\in\mathbb{Z}} h(n)\phi(u-n)} \tag{161}$$

And the $h(n)$ are obtained using the orthogonality of the basis:

$$h(n) = \langle \frac{1}{\sqrt{2}}\phi\left(\frac{u}{2}\right), \phi(u-n)\rangle \tag{162}$$

So, if we know ϕ, we know the coefficients $h(n)$. **But what's even more interesting is that we can start with $h(n)$ and construct ϕ.** Let's move to the Fourier domain, which gives us

$$\sqrt{2}\hat{\phi}(2\omega) = \left(\sum_{n\in\mathbb{Z}} h(n)e^{-in\omega}\right)\hat{\phi}(\omega) = \hat{h}(\omega)\hat{\phi}(\omega) \tag{163}$$

where we recognize the Fourier series associated with h. This leads to

$$\sqrt{2}\hat{\phi}(2\omega) = \hat{h}(\omega)\hat{\phi}(\omega) \iff \hat{\phi}(\omega) = \frac{\hat{h}(\omega/2)}{\sqrt{2}}\hat{\phi}(\omega/2) \tag{164}$$

We want to repeat this process, which allows us to write

$$\hat{\phi}(\omega) = \hat{\phi}(2^{-J}\omega)\prod_{p=1}^{J}\frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \tag{165}$$

Now, if we let J tend to infinity, we see that $\hat{\phi}(\omega)$ is **completely determined by the filter $\hat{h}(\omega)$**:

$$\boxed{\hat{\phi}(\omega) = \hat{\phi}(0)\prod_{p=1}^{+\infty}\frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}}} \tag{166}$$

What are the properties of $\hat{h}(\omega)$? Remember that the family $\{\phi(u-n)\}_{n\in\mathbb{Z}}$ forms

[46] In the literature, this is referred to as the *scaling relation.*

an orthonormal basis and is used to construct a multiresolution. All of this leads to the conclusion that the filter $\hat{h}$ must be quite special. In fact, we want to reverse the problem: what are the properties of h that work well so that ϕ then defines a multiresolution? In doing so, we will rediscover properties used in signal theory, namely, *mirror filters*. Here's the theorem:

Theorem 17 *(filter h)*
If ϕ defines a multiresolution (Def. 4), and if we define $h(n)$ as follows

$$h(n) = \langle \frac{1}{\sqrt{2}}\phi\left(\frac{u}{2}\right), \phi(u-n)\rangle = \langle \phi_1, \phi_{0,n}\rangle \tag{167}$$

then the Fourier transform of h satisfies the relations

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega+\pi)|^2 = 2 \quad \textit{and} \quad \hat{h}(0) = \sqrt{2} \tag{168}$$

These properties are illustrated in Figure 40.

Conversely, if $\hat{h}$ satisfies the above relations and if

$$\hat{h}(\omega) > 0 \quad \forall\omega \in [-\pi/2, \pi/2] \tag{169}$$

then

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \tag{170}$$

is the Fourier transform of a function ϕ that defines a multiresolution.

Proof 17. Assuming[47] that $\hat{\phi}(0) \neq 0$, we can easily deduce that $\hat{h}(0) = \sqrt{2}$ from Eq. 164. The first of the two properties in Eq. 168 is the most crucial one and was highlighted by M.J.T. Smith and T.P. Barnwell in the 1980s in the context of filter banks. Recall that we want to end up with an orthonormal basis with $\phi(u)$ translated. Thus, we want

$$\langle \phi(u), \phi(u-n)\rangle = \delta[n] \tag{171}$$

[47] Note that an argument is given later.

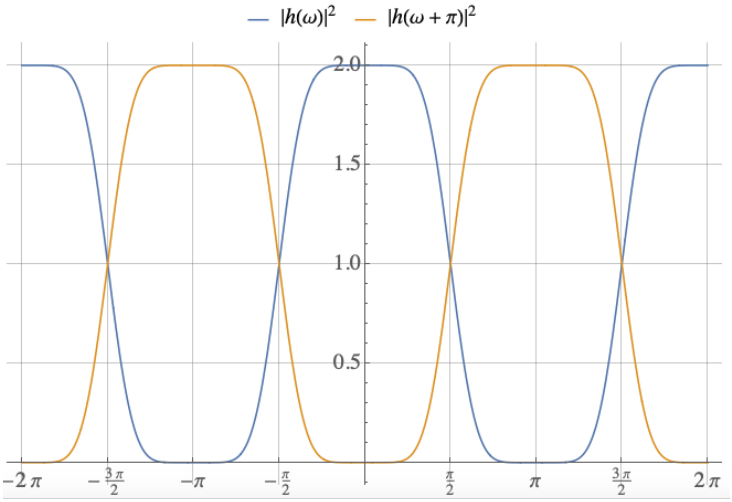

FIGURE 40 – Example of a filter $\hat{h}(\omega)$ that satisfies the relations in Eqs. 168. In this case, it's a filter from a Daubechies multiresolution of order 4. *Note: Depending on the libraries, the filter may or may not include the factor* $\sqrt{2}$.

which, through convolution with $\tilde{\phi}$ (Eq. 158 with $j = 0$), gives

$$(\phi * \tilde{\phi})(n) = \delta[n] \tag{172}$$

Now, let's use Theorem 14 with $T = 1$:

$$\sum_{n\in\mathbb{Z}} x(n)e^{-in\omega} = \sum_{k\in\mathbb{Z}} \hat{x}\left(\omega - 2k\pi\right) \tag{173}$$

For $x = \phi * \tilde{\phi}$, in the case where ϕ is a real function, we have $\hat{x} = |\phi|^2$ and the following relation:

$$\boxed{\sum_{k\in\mathbb{Z}} |\hat{\phi}(\omega - 2k\pi)|^2 = 1} \tag{174}$$

This is the necessary and sufficient condition for having an orthonormal basis. Now, let's use Eq. 164 again. When we replace it, we get

$$\sum_{k\in\mathbb{Z}} |\hat{\phi}(\omega - 2k\pi)|^2 = \frac{1}{2}\sum_{k\in\mathbb{Z}} |\hat{h}(\omega/2 - k\pi)\hat{\phi}(\omega/2 - k\pi)|^2 = 1 \tag{175}$$

Therefore, by separating the even and odd k terms and using the fact that $\hat{h}(\omega)$ is 2π-periodic (i.e., a Fourier series), we can write

$$\begin{aligned}
&\sum_{k\in\mathbb{Z}} |\hat{h}(\omega/2-k\pi)|^2|\hat{\phi}(\omega/2-k\pi)|^2 \\
&\qquad= \sum_{p\in\mathbb{Z}} |\hat{h}(\omega/2-2p\pi)|^2|\hat{\phi}(\omega/2-2p\pi)|^2 + \sum_{p\in\mathbb{Z}} |\hat{h}(\omega/2-(2p+1)\pi)|^2|\hat{\phi}(\omega/2-(2p+1)\pi)|^2 \\
&\qquad= \sum_{p\in\mathbb{Z}} |\hat{h}(\omega/2)|^2|\hat{\phi}(\omega/2-2p\pi)|^2 + \sum_{p\in\mathbb{Z}} |\hat{h}(\omega/2-\pi)|^2|\hat{\phi}(\omega/2-(2p+1)\pi)|^2 \\
&\qquad= |\hat{h}(\omega/2)|^2 \underbrace{\sum_{p\in\mathbb{Z}} |\hat{\phi}(\omega/2-2p\pi)|^2}_{} + |\hat{h}(\omega/2-\pi)|^2 \underbrace{\sum_{p\in\mathbb{Z}} |\hat{\phi}(\omega/2-(2p+1)\pi)|^2}_{}
\end{aligned}$$

Now, using relation 174, we can see that the two sums inside the braces are equal to 1. Therefore,

$$|\hat{h}(\omega/2)|^2 + |\hat{h}(\omega/2-\pi)|^2 = 2$$

which completes the proof because this holds for all ω, and we can use the 2π-periodicity of $\hat{h}$ to introduce the $+\pi$ in the second term.

For the converse, we need to consider what happens at $\omega = 0$. If $\hat{\phi}(0) = 0$, then we have a bandpass filter. But, recall the condition on ψ: as j becomes more and more negative, the spectral band shifts towards high frequencies, depleting the low frequencies. Regarding ϕ, this is not possible because the projection $P_{V_j}x$ must satisfy property (iii) of multiresolution (Def. 4), which states that $P_{V_j}x$ must converge to x for every element of $L^2(\mathbb{R})$. This is a contradiction when depleting low frequencies. Thus,

$$\hat{\phi}(0) \neq 0$$

The complete proof of the converse is in the notes that S. Mallat attaches to his course. It starts with the relations on $\hat{h}(\omega)$ to demonstrate that the product in Eq. 170 makes sense and indeed yields a function ϕ that has property 174 to provide an orthonormal basis. ■

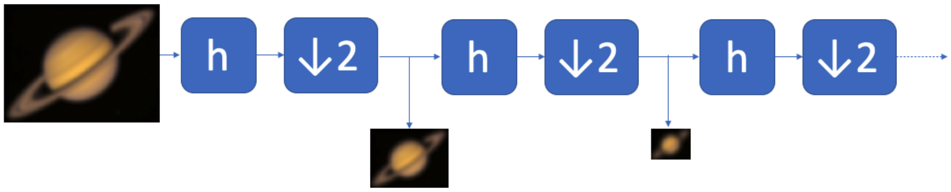

FIGURE 41 – Burt's Algorithm: cascade of a low-pass filter $\hat{h}(\omega)$ to prevent aliasing, followed by downsampling by a factor of 2.

7.3 Filter Bank Algorithms (I)

Once we have a function $\hat{h}(\omega)$ that satisfies the relations 174, or its real-space version using $h(n)$, i.e.,

$$\hat{h}(\omega) = \sum_{n \in \mathbb{Z}} h(n) e^{-in\omega} \tag{176}$$

we would like to understand how this can be useful in practice. In fact, algorithms existed in image processing, notably Peter Burt's algorithm [48]. It involved cascading a low-pass filter h to prevent aliasing (Sec. 6.4), followed by downsampling by a factor of 2 (Fig. 41). The challenge at that time was to perform these operations in real-time, which required using small-sized filters for fast convolutions. Although the filters used did not satisfy the above-mentioned properties, the cascade of operations was in place.

Now, the orthogonal projection of x onto V_j is given by Theorem 16:

$$x(u) = \sum_{n \in \mathbb{Z}} a_j[n] \phi_{j,n}(u) \tag{177}$$

So, we transition from the continuous variable u to the discrete variable n, which means moving from $x(u)$ to $a_j[n]$

$$a_j[n] = \langle x, \phi_{j,n} \rangle = (x * \tilde{\phi})(2^j n) \tag{178}$$

and we are now dealing with **discrete sequences** only. What is the relationship between $a_j[n]$ and $a_{j+1}[n]$? The relationship becomes clear when we have the nested structure of

[48] Burt-Adelson pyramid image processing.

spaces V_j: knowing the projection of x onto V_j, we can project it onto space V_{j+1} and so on. To obtain the projection into V_{j+1}, we need to know the inner products with $\phi_{j+1,n}$. Now,

$$\phi_{j+1}(u) = \sum_n \langle \phi_{j+1}, \phi_{j,n} \rangle \phi_{j,n}(u) \tag{179}$$

with

$$\begin{aligned} \langle \phi_{j+1}, \phi_{j,n} \rangle &= \int \frac{1}{\sqrt{2^{j+1}}} \phi\left(\frac{u}{2^{j+1}}\right) \frac{1}{\sqrt{2^j}} \phi\left(\frac{u-2^j n}{2^j}\right)\, du \\ &= \frac{1}{\sqrt{2}} \int \phi\left(\frac{u}{2}\right) \phi(u-n)\, du \\ &= h(n) \end{aligned} \tag{180}$$

where the last equality is obtained using Theorem 17. Thus, we arrive at the relation

$$\boxed{\phi_{j+1}(u) = \sum_{n\in\mathbb{Z}} h(n) \phi_{j,n}(u)} \tag{181}$$

What about $\phi_{j+1,n}(u)$? It follows that

$$\begin{aligned} \phi_{j+1,p}(u) = \phi_{j+1}(u - 2^{j+1}p) &= \sum_{n\in\mathbb{Z}} h(n) \phi_{j,n}(u - 2^{j+1}p) \\ &= \sum_{n\in\mathbb{Z}} h(n) \frac{1}{\sqrt{2^j}} \phi\left(\frac{u - 2^j(2p) - 2^j n}{2^j}\right) \end{aligned} \tag{182}$$

and thus, we obtain the relation

$$\boxed{\phi_{j+1,p}(u) = \sum_{n\in\mathbb{Z}} h(n) \phi_{j,n+2p}(u)} \tag{183}$$

We can now calculate $a_{j+1}[p]$:

$$a_{j+1}[p] = \langle x, \phi_{j+1,p} \rangle = \sum_{n\in\mathbb{Z}} h(n) \langle x, \phi_{j,n+2p} \rangle \tag{184}$$

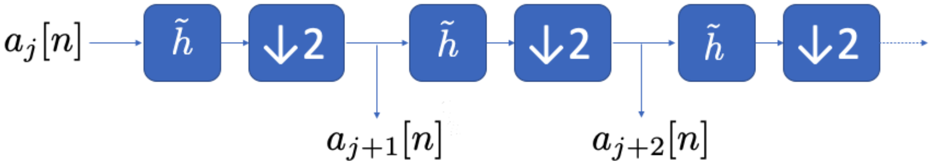

Figure 42 – Algorithm for transitioning from $a_j[n]$ to $a_{j+1}[n]$: cascade of a discrete low-pass filter $\tilde{h}$ followed by downsampling by a factor of 2. Similarities and differences with Burt's algorithm in Figure 41 can be noted.

and thus obtain

$$\boxed{a_{j+1}[p] = \sum_{n\in\mathbb{Z}} h(n)a_j[n+2p] = \sum_{n\in\mathbb{Z}} h(n-2p)a_j[n] = (a_j * \tilde{h})(2p)} \tag{185}$$

where we have highlighted the convolution with the filter $\tilde{h}[n] = h[-n]$ (the counterpart of $\tilde{\phi}$) taken at $2p$, meaning we indeed have a **cascade of filtering and downsampling** (Fig. 42).

7.3.1 Example with Haar Multiresolution

Let's revisit the indicator function ϕ of $[0,1]$, which was our first example of multiresolution (Sec. 7.1.2). The translated functions $\phi(u-n)$ are, therefore, indicators:

$$\phi(u-n) = \mathbf{1}_{[n,n+1]}(u) \tag{186}$$

Similarly, the function ϕ scaled by a factor of 2 is also an indicator:

$$\frac{1}{\sqrt{2}}\phi\left(\frac{u}{2}\right) = \frac{1}{\sqrt{2}}\mathbf{1}_{[0,2]}(u) \tag{187}$$

Thus, using Theorem 17, we define the discrete filter h as the inner product between $\phi(u-n)$ and $\frac{1}{\sqrt{2}}\phi\left(\frac{u}{2}\right)$, which gives

$$h[n] = \begin{cases} 1/\sqrt{2} & n = 0, 1 \\ 0 & \text{elsewhere} \end{cases} \tag{188}$$

Now, according to Theorem 16, consider the signal $x(u)$ with u as a continuous variable. In practice, this signal is sampled on a fixed grid denoted as V_0, and we have the coefficients $a_0[n] = x(n)$. Next, the algorithm provides the coefficients $a_1[p]$:

$$a_1[p] = \frac{a_0[2p] + a_0[2p+1]}{\sqrt{2}} \tag{189}$$

and the projection onto V_1 of the signal becomes

$$P_{V_1}x(u) = \sum_p a_1[p]\phi_1(u-p) = \sum_p \frac{a_0[2p] + a_0[2p+1]}{2}\phi\left(\frac{u}{2} - p\right) \tag{190}$$

Note that we ultimately average the coefficients to transition from one grid to another. An illustration of this algorithm is shown in Figure 43. What will change in the future is the support of h and its values, but the algorithm will remain the same.

7.4 Connection with Wavelet Bases

The filter bank algorithm studied in the previous section provides successive averaging, but this is not how we can construct sparse representations where we aim to obtain zeros instead. Note that when transitioning from V_j to V_{j+1} and reducing information, we lose the details of the signal present in V_j due to averaging. If we want to highlight these details, we should project not into V_{j+1} but into the complementary space[49] W_{j+1} (Fig. 44):

$$V_j = W_{j+1} \oplus V_{j+1} \tag{191}$$

[49] Note for readers from the 2018 Course, the scales are indexed by $-j$, meaning $s = 2^{-j}$.

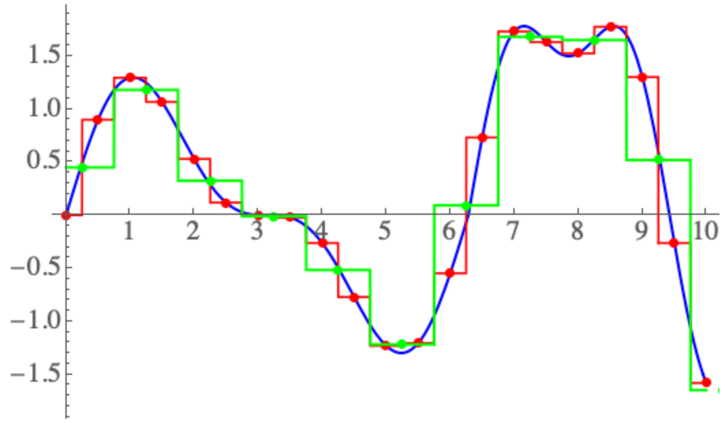

FIGURE 43 – Illustration of the sampling algorithm with the Haar function $\phi(u)$: first, we sample the signal to obtain an approximation (red) on grid V_0, then in a second step, through filtering and downsampling, we obtain an approximation on a grid 2 times larger, V_1.

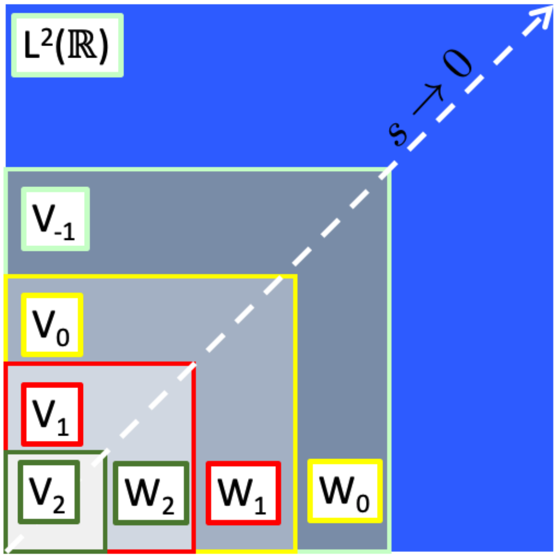

FIGURE 44 – Spaces V_j and W_j connected by the complementarity relation Eq. 191. This complements Figure 38.

We know an orthonormal basis in V_j and V_{j+1}, so we need to construct an orthogonal basis in W_{j+1}. However, note that the previous relation calls for recursion:

$$V_j = \bigoplus_{p=j+1}^{J} W_p \oplus V_J \tag{192}$$

Now, if $J \to +\infty$, V_J tends towards the empty set, while if $j \to -\infty$, then V_j tends towards the entire space $L^2(\mathbb{R})$. In a sense,

$$L^2(\mathbb{R}) = \bigoplus_{j=-\infty}^{+\infty} W_j \tag{193}$$

and all the spaces W_j are orthogonal to each other. Thus, obtaining an orthogonal basis of $L^2(\mathbb{R})$ is done by grouping the orthogonal bases of all the spaces W_j. This is where the wavelet ψ comes into play to connect the W_j sets with each other.

Theorem 18 *(filter g)*
Let $g(n)$ be a filter obtained from the filter h (Th. 17) as follows:

$$g(n) = (-1)^{1-n} h(1-n) \tag{194}$$

Also, let $\psi(u)$ be derived from this filter g and the function ϕ (which is also derived from h) as follows:

$$\frac{1}{\sqrt{2}}\psi\left(\frac{u}{2}\right) = \sum_{n\in\mathbb{Z}} g(n)\phi(u-n) \tag{195}$$

Then, for all j, the family

$$\left\{\psi_{j,n}(u) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{u-2^j n}{\sqrt{2^j}}\right)\right\}_{n\in\mathbb{Z}} \tag{196}$$

is an orthonormal basis of W_j. And the family $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$ is an orthonormal basis of $L^2(\mathbb{R})$.

This provides a generalization of Haar from the given filter h. It is remarkable that it is possible to construct wavelet bases that go beyond the scope of this theorem, but such wavelets are pathological and have very slow decay, making them unattractive and impractical for use. Therefore, in practice, we can view this theorem as a necessary and sufficient condition for obtaining orthonormal wavelet bases.

Proof 18. First, if $\psi(u) \in W_1 \subset V_0$, the orthogonal complement of V_1 in V_0, and since V_0 has an orthonormal basis $\{\phi_n\}_{n\in\mathbb{Z}}$, we can decompose $\psi(u/2)$ as follows:

$$\frac{1}{\sqrt{2}}\psi\left(\frac{u}{2}\right) = \sum_{n\in\mathbb{Z}} g(n)\phi(u-n) \tag{197}$$

with $g(n)$ unknown for now. But we know that

$$g(n) = \langle \frac{1}{\sqrt{2}}\psi\left(\frac{u}{2}\right), \phi(u-n)\rangle \tag{198}$$

We also want an orthonormal basis[50] of W_0 from $\{\psi_n\}_{n\in\mathbb{Z}}$, which means

$$\langle\psi(u), \psi(u-n)\rangle = \langle\psi, \psi_n\rangle = \delta[n] \Leftrightarrow (\psi * \tilde{\psi})(n) = \delta[n] \tag{199}$$

So, following the same reasoning as for ϕ (Eq. 174), we obtain the relation:

$$\boxed{\sum_{k\in\mathbb{Z}} |\hat{\psi}(\omega - 2k\pi)|^2 = 1} \tag{200}$$

But we also want W_0, the complement of V_0 in V_{-1}, to be orthogonal to V_0, so the family $\{\psi_n\}_{n\in\mathbb{Z}}$ must be orthogonal to the family $\{\phi_n\}_{n\in\mathbb{Z}}$. This additional constraint in the Fourier domain results in the following relation:

$$\boxed{\sum_{k\in\mathbb{Z}} \hat{\phi}^*(\omega - 2k\pi)\hat{\psi}(\omega - 2k\pi) = 0} \tag{201}$$

Now, these two relations 200 and 201 are equivalent to the following two relations:

$$|\hat{g}(\omega)|^2 + |\hat{g}(\omega+\pi)|^2 = 2 \tag{202}$$

$$\hat{g}(\omega)\hat{h}^*(\omega) + \hat{g}(\omega+\pi)\hat{h}^*(\omega+\pi) = 0 \tag{203}$$

[50] Note that we could write that $\{\psi_{1,n}\}_{n\in\mathbb{Z}}$ form an orthonormal basis of W_1.

Finally, one solution for $\hat{g}(\omega)$ is to choose:

$$\boxed{\hat{g}(\omega) = e^{-i\omega}\hat{h}^*(\omega+\pi)} \tag{204}$$

In this case, it follows that:

$$\begin{aligned} \hat{g}(\omega) &= e^{-i\omega}\sum_{n\in\mathbb{Z}} h^*(n)e^{i(\omega+\pi)n} \\ &= \sum_{n\in\mathbb{Z}}(-1)^n h^*(n)e^{-i\omega(1-n)} \\ &= \sum_{m\in\mathbb{Z}}(-1)^{1-m} h^*(1-m)e^{-i\omega m} \end{aligned} \tag{205}$$

Hence, the identification of $g(n)$ yields:

$$g(n) = (-1)^{1-n}h^*(1-n) \tag{206}$$

Since $\phi(u)$ is real, it follows that $h(n)$ is also real, which concludes the proof. ■

8. Lecture 3 Mar.

We saw in the last session how, starting from a function $\phi(x)$ or its filter $h[n]$ (low-pass), we can construct low-dimensional approximations at successive scales of the signal $x(u)$ by projecting them onto the spaces V_j:

$$P_{V_j}x(u) = \sum_{n\in\mathbb{Z}} a_j[n]\phi_{j,n}(u) \tag{207}$$

where $a_j[n] = \langle x, \phi_{j,n}\rangle$. Similarly, to introduce details that are lost when changing the scale through low-frequency approximation, we can refine using the complementarity between spaces V_j and W_j as follows (Fig. 44):

$$V_{j-1} = V_j \oplus W_j \tag{208}$$

and thus

$$P_{V_{j-1}}x = P_{V_j}x \oplus P_{W_j}x \tag{209}$$

We introduced a wavelet ψ and its associated band-pass filter $g[n]$. We also obtained the relationships that filters h and g have (Th. 17).

8.1 Some Examples of Orthonormal Bases

In Sections 6.3.3 and 7.1.2, we saw some examples of wavelets. What about the associated filters?

For the **Haar** wavelet, this was done in Section 7.3.1 with a numerical application. As a reminder[51]:

$$h^{Haar}[n] = \begin{cases} 1/\sqrt{2} & n = 0, 1 \\ 0 & \text{elsewhere} \end{cases} \tag{210}$$

which yields for the filter of $\phi(u) = \mathbf{1}_{[0,1]}(u)$:

$$g^{Haar}[n] = \begin{cases} g[0] = -h[1],\ g[1] = h[0] & \\ 0 & \text{elsewhere} \end{cases} \tag{211}$$

Thus, if $h[n]$ corresponds to an average of samples two by two, $g[n]$ corresponds to their difference. Furthermore:

$$\psi(u) = \sqrt{2} \sum_{n \in \mathbb{Z}} g[n]\phi(2u - n) \tag{212}$$

so

$$\begin{aligned} \psi^{Haar}(u) &= -\mathbf{1}_{[0,1]}(2u) + \mathbf{1}_{[0,1]}(2u - 1) \\ &= -\mathbf{1}_{[0,1/2]}(u) + \mathbf{1}_{[1/2,1]}(u) \end{aligned} \tag{213}$$

which corresponds to the sign changes shown in Figure 32.

[51] Note that the normalization $\hat{h}(0) = \sqrt{2} = \sum_{n \in \mathbb{Z}} h[n]$ might be surprising. Additionally, conventions may vary depending on the libraries used.

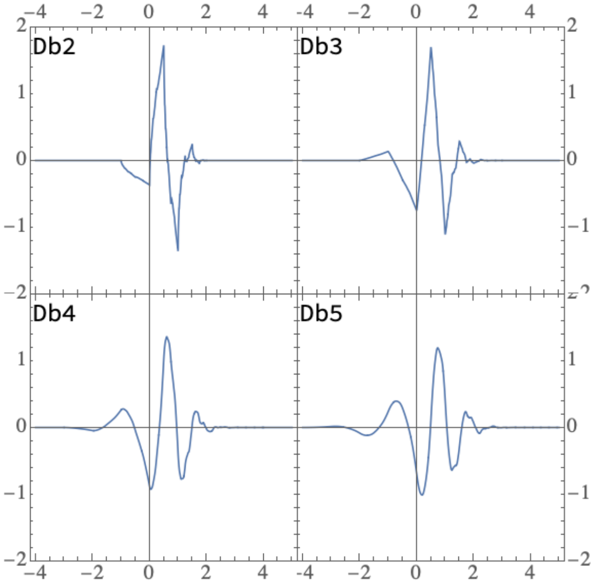

FIGURE 45 – I. Daubechies "Dbm" wavelet illustrations with m the number of zero moments

As a second example, we can consider the **Shannon** wavelet. The Fourier spectrum of ψ is an ideal band-pass (Fig. 33). We can obtain it within the framework of multiresolutions by defining the ideal low-pass 2π-periodic function as:

$$\hat{h}^{Shan}(\omega) = \mathbf{1}_{[-\pi/2,+\pi/2]}(\omega) \tag{214}$$

From which we define the 2π-periodic function $\hat{g}(\omega) = e^{-i\omega}h^*(\omega+\pi)$, which becomes an ideal band-pass:

$$\hat{g}^{Shan}(\omega) = e^{-i\omega}\left\{\mathbf{1}_{[-\pi,-\pi/2]}(\omega) + \mathbf{1}_{[\pi/2,\pi]}(\omega)\right\} \tag{215}$$

Thus, the filter $\hat{h}^{Shan}$ selects low frequencies between $-\pi/2$ and $\pi/2$, while the filter $\hat{g}^{Shan}$ excludes them.

Now, the question arises as to what wavelets, or filters, or multiresolutions are interesting, considering the issues we've seen with Haar and Shannon wavelets, which are discontinuous either in the real space or in the Fourier space. This question has been addressed both by the Signal Processing community and the mathematical community, particularly by the Belgian-American mathematician **Ingrid Daubechies**, who studied how to obtain "optimal" wavelets.

What we want are wavelets with the **smallest possible support**. We also want to be able to assess the Lipschitz-α regularity of signals $x(u)$ by comparing them to a polynomial representation (Th. 3). Therefore, we also need the wavelet ψ to be "transparent" regarding polynomials of a certain degree m, which means that the wavelet has m **vanishing moments** (Eq. 126). As per Theorem 12, we know that the decay of wavelet coefficients $|W_x(v,s)|$ provides information about the regularity order α of x, i.e., $|\langle x, \psi_j \rangle| < C 2^{j(\alpha+1/2)}$.

So, we have two constraints to satisfy: is it possible? The answer is yes. If ψ has m vanishing moments and a compact support, then $\hat{\psi}(\omega) = O(\omega^m)$ near 0 [52]. In other words, imposing vanishing moments on ψ imposes how the band-pass filter $\hat{\psi}(\omega)$ decays near 0. Moreover, $\hat{g}(\omega) = O(\omega^m)$ because $\sqrt{2}\hat{\psi}(2\omega) = \hat{g}(\omega)\hat{\phi}(\omega)$, and $\hat{\phi}(\omega)$ doesn't vanish at 0. Interestingly, the converse is also true, which can be summarized as follows:

> **Property 1**
>
> $$\psi(u) \textit{ has } m \textit{ vanishing moments + compact support} \Leftrightarrow \hat{\psi}(\omega) = O(\omega^m) \Leftrightarrow \hat{g}(\omega) = O(\omega^m)$$

Therefore, to impose vanishing moments on $\psi(u)$, we will either construct or verify that the filter $\hat{g}(\omega)$ has the properties of decaying near 0. We will use the following property as well:

> **Property 2**
>
> $$\psi(u) \textit{ has a compact support} \Leftrightarrow h[n] \textit{ has a compact support}$$

This property can be deduced from the following reasoning: If $\phi(u)$ has compact support, according to the definition of $h[n]$ as the inner product between $\phi(u)$ and $\phi(2u-n)$, we conclude that there are only a finite number of n for which $h[n] \neq 0$. Then, from the relation between $g[n]$ and $h[n]$, we deduce that $g[n]$ is also non-zero for a finite number

[52] Note: with a bit of calculus and using the derivatives of the Fourier transform of ψ, we can show that $\partial^{(n)}\hat{\psi}(\omega) = 0$ for all $n \leq m$, and then use Taylor's expansion of $\hat{\psi}(\omega)$.

of n. Finally, the relation between $\psi(u)$, $\phi(2u-n)$, and $g[n]$ ensures that ψ has compact support. Conversely, if $h[n]$ does not have compact support, then $\phi(u/2)$ and $\phi(u-n)$ have a non-zero overlap, so $\phi(u)$ cannot have compact support.

In fact, what Ingrid Daubechies demonstrated is that we cannot satisfy all the properties we would like at the same time. In particular, **if we want to increase the number of vanishing moments, then the support of ψ (and ϕ) must grow**. Thus, satisfying the spatial localization of ψ and the regularity of the filter $\hat{\psi}$ near 0 are in balance, and there is an optimization to be done.

> **Property 3** *$\psi(u)$ has m vanishing moments and defines an orthonormal wavelet basis $\Rightarrow$ Support$(\psi) \geq 2m-1$, and the filters h and g have a size of $2m$.*

(NDJE: The case $m = 1$ yields the Haar wavelet, which indeed has only one vanishing moment, namely, that its integral is zero, but it cannot make a monomial of degree 1 vanish.)

Several wavelets by I. Daubechies are shown in Figure 45, with "Dbm" indicating the number of vanishing moments (for Haar/"Db1", the wavelet has only one vanishing moment). It can be observed that as m increases, the wavelet becomes smoother, in addition to having its support increased.)

8.2 Filter Bank Algorithms (II): DWT/IDWT

In Section 7.3, we developed a cascade of filtering-downsampling algorithms using the low-pass filter $h[n]$ (Fig. 41) to obtain successive low-dimensional approximations $P_{V_j}x$. The idea now is to introduce[53] the details of x by applying the band-pass filter $g[n]$ to obtain $P_{W_j}x$. A generic decomposition example that will serve as a reference is shown in Figure 46.

So, the idea is to start with sampling $a_L[n]$ of the signal $x(u)$, which constitutes the approximation at a certain reference scale L (the number of samples is $N_s = 2^L$, and the

[53] See also Course 2018 Sec. 6.6.0.4.

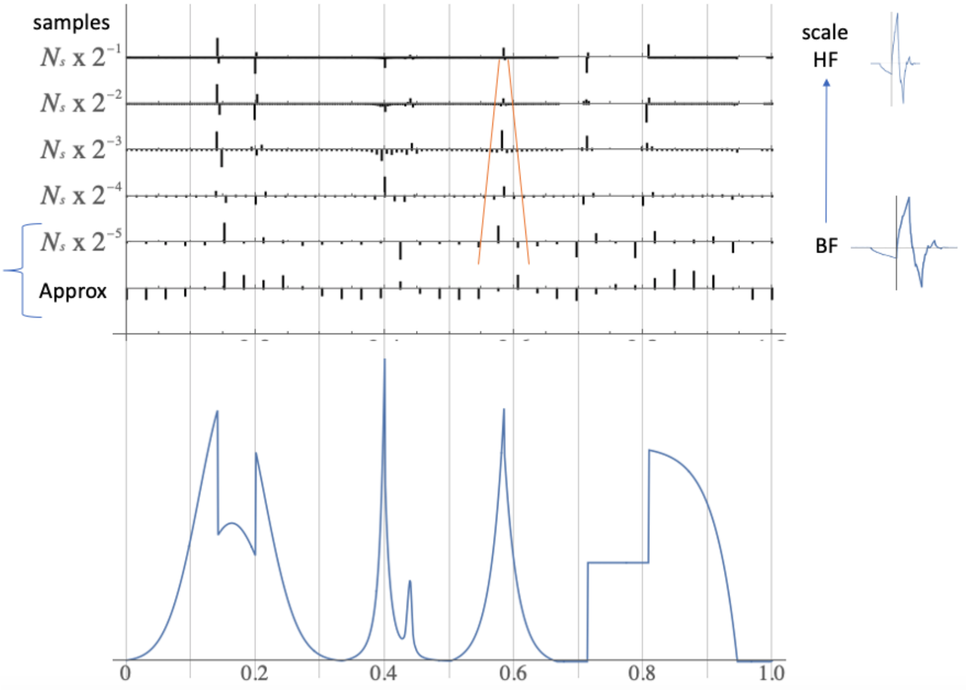

FIGURE 46 – From bottom to top: the sampled function yielding $2^S = 1024$ samples ($S = 10$), then by successive decomposition with the "Db2" wavelet, we obtain low-frequency approximation coefficients numbering $2^{S-5} = 32$ and detail coefficients at the same scale. We also have detail coefficients of increasingly higher frequencies, ending with those numbering $2^{S-1} = 512$. Note that most of the detail coefficients are nearly zero, and the most important detail coefficients concentrate at the discontinuities of the original function in the cone of Theorem 12 (orange). The vertical scales are not the same for each set of coefficients. *Note: S. Mallat's presentation on the projector is in reverse order for the arrangement of detail coefficients, going from top to bottom, from low to high frequencies.*

sampling interval $\Delta_u \propto 1/N_s$), i.e.,

$$P_{V_L}x(u) = \sum_n a_L[n]\ \phi_{L,n}(u) \tag{216}$$

From these samples, we can reconstruct wavelet coefficients at larger spatial scales $2\Delta u$, $2^2\Delta u$, up to a limit scale where the size of the approximation's support is of the same order of magnitude as the sampled function (accounting for edge effects). How do we proceed?

If we know $P_{V_{j-1}}x$, we know that

$$\begin{aligned} P_{V_{j-1}}x &= P_{V_j}x + P_{W_j}x \\ \Rightarrow \sum_n a_{j-1}[n]\phi_{j-1,n}(u) &= \sum_n a_j[n]\phi_{j,n}(u) + \sum_n d_j[n]\psi_{j,n}(u) \end{aligned} \tag{217}$$

Now, we know the decomposition of $\phi_{j,n}$ based on $\phi_{j-1,m}$ (Eq. 183), and we get

$$\langle \phi_{j-1,m}, \phi_{j,n} \rangle = h[m-2n] \tag{218}$$

Since $\phi_{j,n}$ and $\psi_{j,m}$ are orthogonal, we can proceed with the inner product with $\phi_{j,n}$ (while anonymizing the index n)

$$\boxed{a_j[n] = \sum_{m\in\mathbb{Z}} h[m-2n]a_{j-1}[m] = (a * \tilde{h})[2n]} \tag{219}$$

where $\tilde{h}[n] = h[-n]$. Similarly, we can show, based on the same framework developed in Section 7.3, that

$$\boxed{\psi_{j+1,p}(u) = \sum_{n\in\mathbb{Z}} g[n]\phi_{j,n+2p}(u)} \tag{220}$$

and, therefore, the detail coefficients are derived from a_{j-1} as follows:

$$\boxed{d_j[n] = \sum_{m\in\mathbb{Z}} g[m-2n]a_{j-1}[m] = (a * \tilde{g})[2n]} \tag{221}$$

The elementary cell of the DWT algorithm is presented in Figure 47, and the DWT (Discrete Wavelet Transform) algorithm is shown in Figure 48.

So, from an algorithmic perspective, the transition from a_{j-1} to (a_j, d_j) is done using

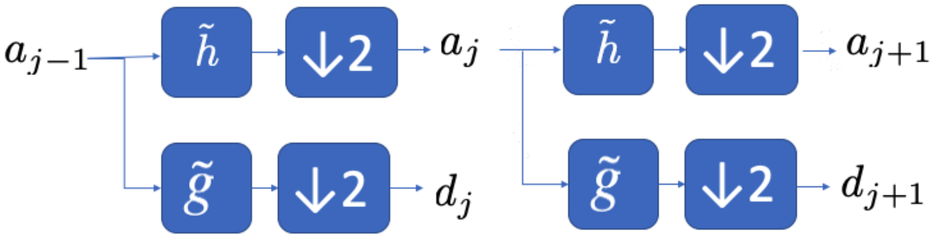

FIGURE 47 – From the approximation coefficients a_{j-1}, we obtain the approximation coefficients a_j and detail coefficients d_j (Eqs. 219, 221). Then, we can cascade the algorithm starting from a_j, and so on.

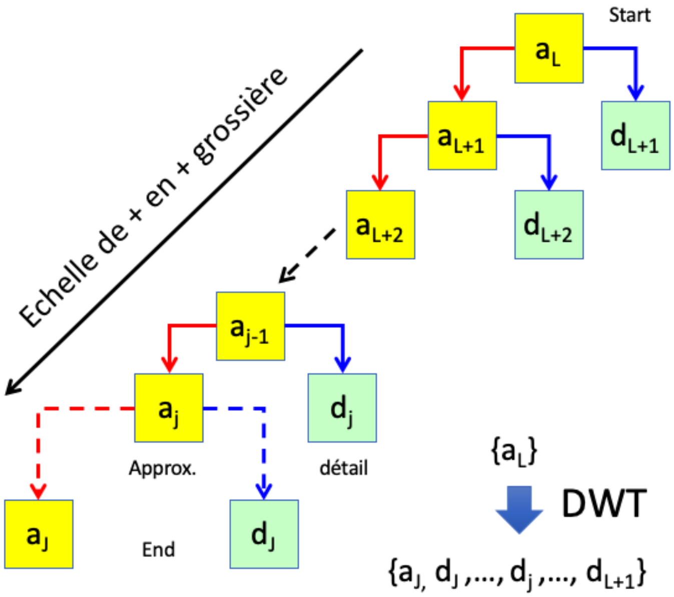

FIGURE 48 – Diagram of a complete wavelet decomposition (Discrete Wavelet Transform).

the filters h and g, which makes finite-support Daubechies filters attractive. However, it's important to note that underlying this, we calculate inner products with wavelets that determine all the properties of these coefficients (e.g., sparsity). What is the computational cost of the DWT transformation? At each cell of the decomposition from a_{j-1} to (a_j, d_j), we have a constant number of operations if the filters have finite support ($2m$), namely, $2m$ multiplications and $2m$ additions per coefficient. Furthermore, this number of coefficients is halved at each stage. Therefore, if we start with N samples at scale L and decompose down to the coarsest scale J, the number of operations is equal to

$$\sum_{j=1}^{L-J} N2^{-j}(4m) = 4mN(1 - 2^{J-L}) \leq 4mN \tag{222}$$

In other words, **the number of operations for the DWT is linear in** N, making it faster than the FFT, which is $O(N \log N)$. Additionally, the smaller the support of the filters, the faster the DWT.

The algorithm is invertible (Inverse Discrete Wavelet Transform, or IDWT), and the nested structure and orthonormal bases provide the synthesis formulas. Starting from Equation 217 and the inner products

$$\langle \phi_{j-1,m}, \phi_{j,n} \rangle = h[m - 2n] \qquad\qquad \langle \phi_{j-1,m}, \psi_{j,n} \rangle = g[m - 2n] \tag{223}$$

we obtain the following relation, which allows us to construct a finer-scale (or higher-frequency) approximation at scale $j - 1$:

$$\boxed{a_{j-1}[n] = \sum_{m \in \mathbb{Z}} (a_j[m]h[n - 2m] + d_j[m]g[n - 2m])} \tag{224}$$

To highlight a convolution operation, due to the $2m$ in the filters, we redefine a_j and d_j by inserting zeros as follows:

$$\check{a}_{j-1}[n'] = \begin{cases} a_{j-1}[n] & n' = 2n \\ 0 & n' = 2n + 1 \end{cases} \tag{225}$$

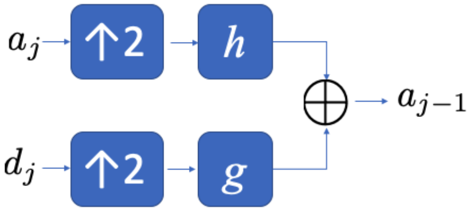

FIGURE 49 – From the coefficients a_j and d_j, we can reconstruct the coefficients a_{j-1} using the formulas in Eqs. 224, 226, with the latter highlighting the doubling of the sizes of coefficients a_j and d_j by inserting zeros before filtering.

Similarly for $\check{d}_{j-1}$. Thus, we can write

$$\boxed{a_{j-1}[n] = (\breve{a}_j * h)[n] + (\check{d}_j * g)[n]} \tag{226}$$

The diagram of the elementary cell of the IDWT algorithm is given in Figure 49. Now, by iterating the process, we realize that to reconstruct the approximation of a certain order, we need a_J and d_J, and only the detail coefficients d_{J-1}, d_{J-2}, and so on. The coefficients a_{J-1}, a_{J-2}, etc., are only intermediate calculations. The complete IDWT algorithm is presented in Figure 50.

8.3 Signal Approximations: Experimentation

With the coefficients $\{a_J, d_J, d_{J-1}, \ldots, d_{L+1}\}$, we can construct various types of signal approximations for x. For example, we can eliminate all the high-frequency coefficients (d_j) to obtain **low-frequency linear approximations** $P_{V_j}x$ with coefficients a_j. An example is shown in Figure 51 using the function from Figure 46.

However, if we zoom in on one of these linear approximations (Fig. 52), we notice that errors occur at discontinuities because we have smoothed the signal. We also observe small residual oscillations, which are caused by the Gibbs phenomenon[54].

[54] The "Gibbs phenomenon" (named after Josiah Willard Gibbs, physicist, 1839-1903): it is a phenomenon of non-uniform convergence of Fourier series, first demonstrated by Henry Wilbraham in 1848, then discussed by Albert Michelson (Nobel laureate in Physics, 1852-1931) in 1898 and later by Gibbs in the journal *Nature*. By the way, Michelson's machine (https://www.youtube.com/watch?v=

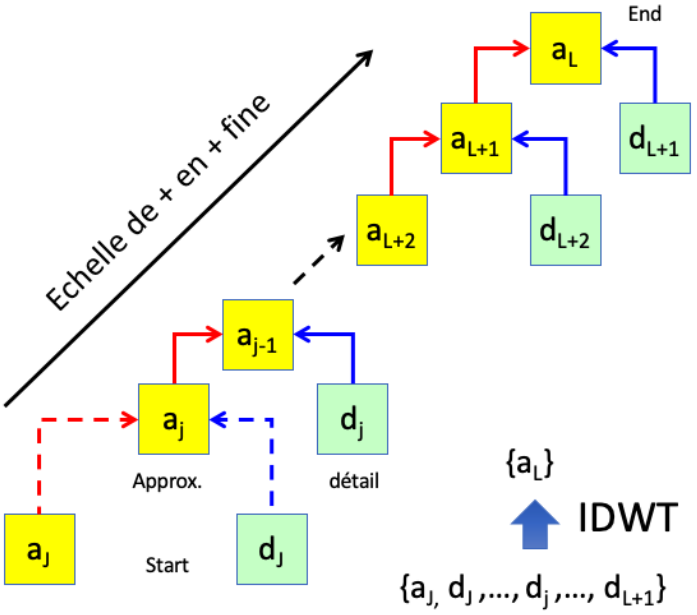

Figure 50 – Diagram of a complete wavelet synthesis (Inverse Discrete Wavelet Transform).

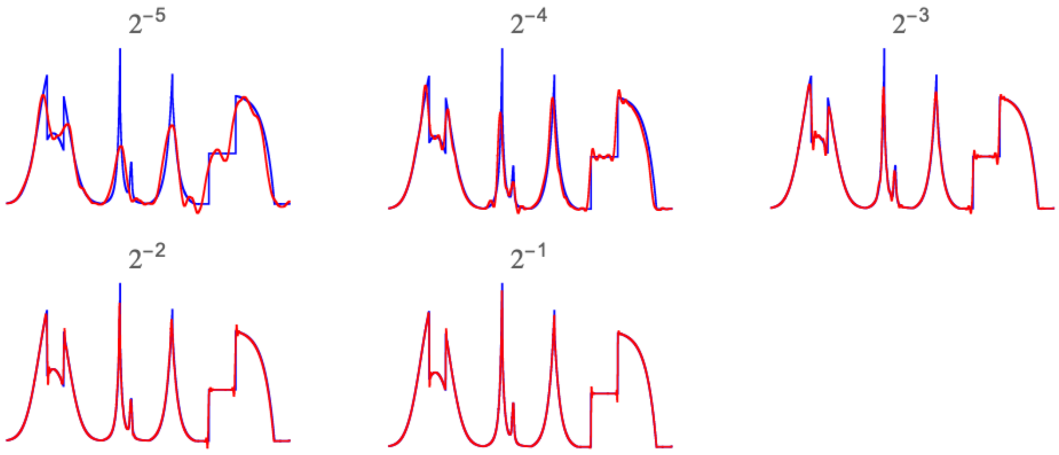

Figure 51 – Approximations of the x signal (Fig. 46) of the type $P_{V_j}x$ obtained with different coefficients a_j from $j = J$ to $j = L + 1$.

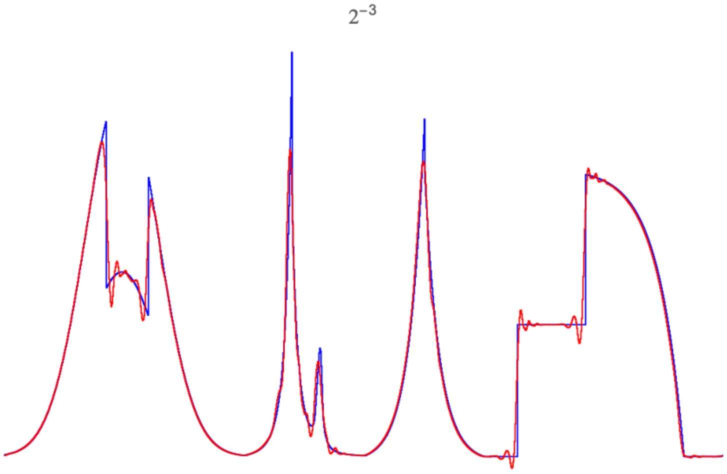

Figure 52 – Extract from Figure 51 to show that errors in low-frequency linear approximation occur at discontinuities, resulting in the Gibbs phenomenon.

How can we do better? As we recall from Section 5.1, we know that to obtain a high-quality **non-linear approximation**, we need to be able to **adapt to the underlying function/signal**, for example, by keeping only **the most significant decomposition coefficients**. For instance, we can take the DWT decomposition that gives the linear approximation like the one labeled "2^{-5}" in Figure 51, corresponding to $J = 5$, and consider only the most important coefficients from the list $\{a_J, d_J, d_{J-1}, \dots, d_{L+1}\}$ to have a total of approximately 128 coefficients, which is about the same number as in the "2^{-3}" linear approximation (Fig. 52). The result is shown in Figure 53, demonstrating a significant improvement in the approximation (note that by increasing to 150 coefficients, the small Gibbs oscillations are no longer visible). It's important to remember that we start with 1024 samples of the original function/signal. In terms of approximation quality with the same number of coefficients, we have:

$$\frac{\|f - f_{\text{app. lin.}}\|^2}{\|f\|^2} = 1.06\ 10^{-2} \qquad\qquad \frac{\|f - f_{\text{app. n-lin.}}\|^2}{\|f\|^2} = 6.95\ 10^{-4}$$

NAsM30MAHLg&feature=em-comments) could not demonstrate this phenomenon, despite the lingering legend. It was only in 1906 that the American mathematician Maxime Bôcher (1867-1918) provided a clear explanation and gave it the name "Gibbs phenomenon", often also referred to as Wilbraham-Gibbs, especially due to the eponymous constant. Subsequently, the definition of this type of low-frequency linear approximation error was extended.

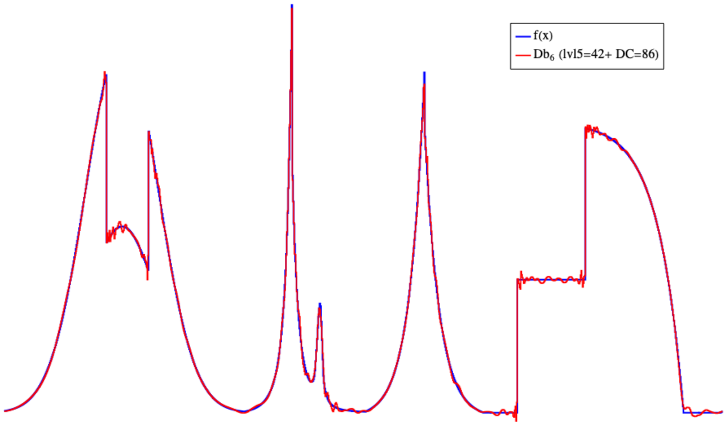

FIGURE 53 – In dark blue, the underlying function/signal, and in red, the non-linear approximation obtained using the 128 most significant coefficients from the DWT decomposition. This non-linear approximation is compared to the linear approximation in Figure 52, which has approximately the same number of coefficients.

So, we were able to obtain a sparse representation of the signal $x(u)$ and achieve a good non-linear approximation that surpasses the linear approximation by more than a factor of 10.

8.4 2D Wavelets

Image processing with wavelets can be done as effectively as in 1D, with one subtlety. Two examples are shown in Figure 54. Why do we need 3 wavelets ψ? In 2D, the variable u has two components (u_1, u_2). One might be tempted to use a separable product with the 1D basis $\{\psi_{j,n}\}_{j,n}$: $\{\psi_{j_1,n_1}(u_1)\psi_{j_2,n_2}(u_2)\}_{j_1,n_1,j_2,n_2}$. The major drawback of this approach is that it favors the two directions u_1 and u_2 with separate scales 2^{j_1} and 2^{j_2}, while we want the decomposition to behave the same if the image is rotated. Therefore, we want wavelet supports associated with scale 2^j in both directions. In fact, we need to go back to the original idea of multiresolutions: the image is first approximated on a regular 2D grid, and gradually, we reduce the resolution by subsampling and examine how to incorporate details to move back up the chain in the reverse direction.

We can separate the scales of the variables at a fixed j to obtain a low-frequency ap-

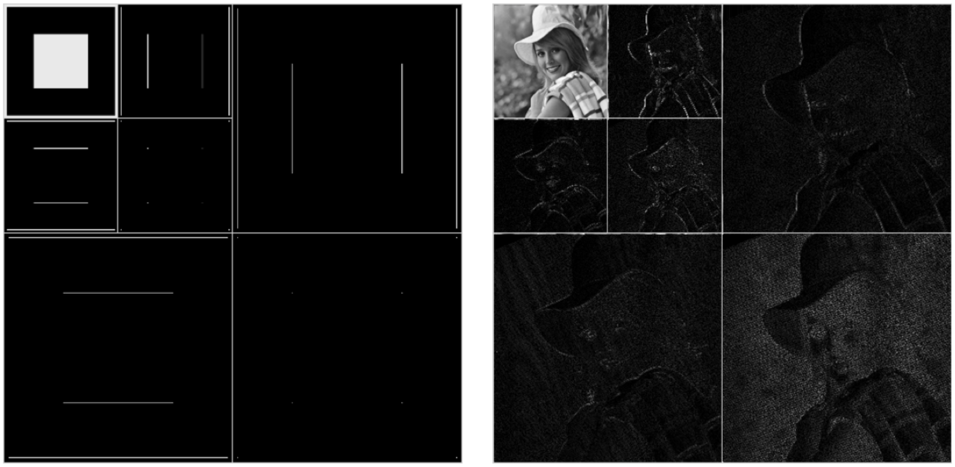

Figure 54 – Examples of wavelet decompositions of two 512×512 images: first, the image is decomposed into a small image constituting the low-frequency approximation of size 256×256 and 3 "detail" sub-images at the same scale sensitive to vertical (top-right), horizontal (bottom-left), and both directions (bottom-right) discontinuities. Then, the low-frequency sub-image is further decomposed, similar to the original image, resulting in 4 more sub-images of size 128×128, and so on. In the top-left corner, we have the coarsest linear approximation sub-image of size 128×128 of the original image. Thus, we have an image pyramid, whereas in 1D, we have a cascade of signals.

proximation using $\{\phi_{j,n_1}(u_1)\phi_{j,n_2}(u_2)\}_{j,n_1,n_2}$, defining a basis for the approximation space V_j on a 2D grid of scale 2^j. Similar to 1D, we can calculate coarser and coarser approximations $P_{V_{j+1}}x$, $P_{V_{j+2}}x$, etc. And just like in 1D, the details will appear in the transformation from $P_{V_j}x$ to $P_{V_{j-1}}x$. To achieve this, we need to decompose V_{j-1} into a direct sum of V_j and a complementary space W_j (Eq. 208). We can look at it from the Fourier frequency perspective (Fig. 55) and realize the necessity of 3 types of wavelets to complete the frequency space between the area covered by the approximation on V_{j-1} (coarser and thus covering a smaller region around the origin) and that on V_j.

Concretely, we define the 3 wavelets that cover the 3 Fourier regions as follows:

$$\begin{cases} \psi^1(u_1,u_2) & = \psi(u_1)\phi(u_2) \\ \psi^2(u_1,u_2) & = \phi(u_1)\psi(u_2) \\ \psi^3(u_1,u_2) & = \psi(u_1)\psi(u_2) \end{cases} \tag{227}$$

As we need to do this at all scales, we dilate/contract and translate these three wavelets as follows:

$$\psi^k_{j,n}(u) = \frac{1}{2^j}\psi^k\left(\frac{u-2^jn}{2^j}\right) \quad u=(u_1,u_2),\ n=(n_1,n_2) \tag{228}$$

What we demonstrate is that **the family** $\left\{\psi^k_{j,n}\right\}_{n\in\mathbb{Z}^2}$ **forms an orthonormal basis for** W_j. Similarly,

$$\phi(u_1,u_2) = \phi(u_1)\phi(u_2) \tag{229}$$

when dilated/contracted and translated also defines a family $\{\phi_{j,n}\}_{n\in\mathbb{Z}^2}$ **which is an orthonormal basis for** V_j. Through recursive complementation, we can obtain **an orthonormal basis for** $L^2(\mathbb{R}^2)$ **with the family** $\left\{\psi^k_{j,n}\right\}_{j\in\mathbb{Z},n\in\mathbb{Z}^2}$.

From an algorithmic perspective, the filtering-subsampling cascade is mostly preserved, but we need to extend the detail coefficients which now cover 3 directions:

$$d_j[n] = (\langle x,\psi^1_{j,n}\rangle, \langle x,\psi^2_{j,n}\rangle, \langle x,\psi^3_{j,n}\rangle) \tag{230}$$

while the low-frequency coefficients (a_j) are formally equal to

$$a_j[n] = \langle x,\phi_{j,n}\rangle \tag{231}$$

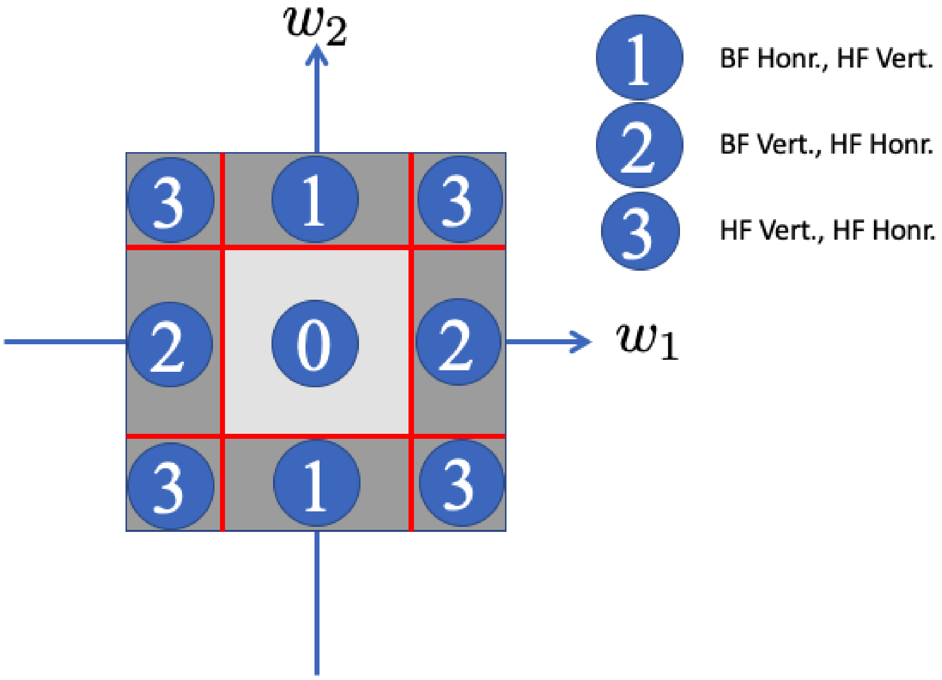

FIGURE 55 – Division of the 2D Fourier plane when transitioning from an approximation V_j (dark gray square) to a coarser approximation V_{j-1} (light gray square): we need to complement this latter area with 3 types of zones numbered 1 to 3, corresponding to frequency pairs that respectively detect: *horizontal edges* or rapidly varying signals along the vertical axis, hence high vertical frequencies; *vertical edges* or rapidly varying signals along the horizontal axis, hence high horizontal frequencies; and finally, edges that exhibit both types of variations.

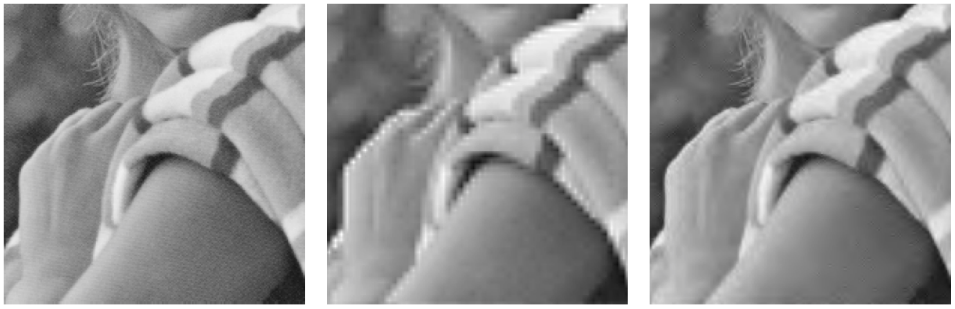

Figure 56 – Inset of an original 512 × 512 image, in the middle a linear approximation that reduces the image size by a factor of 16, and on the right, a non-linear approximation also with 1/16 of the initial coefficients, starting from a decomposition that reduces by a factor of 32 and complements with detail coefficients.

with $\phi_{j,n}$ now acting in both directions (u_1, u_2).

As in 1D, we can obtain much more effective linear and non-linear approximations of an image, as illustrated in Figure 56.

9. Lecture 10 Mar.

9.1 Summary of Concepts Developed in Previous Sessions

Throughout the various sessions, we have explored the RAP triangle, which deals with the issues of *Regularity* that condition low-dimensional **Approximations** at the core of data processing and their connection with *Sparse Representations*. We have shown that there are equivalences between these three concepts, and we can view them from two perspectives: *linear* versus *non-linear*. In the linear approach, we have reviewed that the approximations obtained by projecting onto linear spaces are associated with forms of regularity that are expressed, in particular, through the Fourier basis for translation-invariant problems. In contrast, the non-linear approach leads to approximations that are projections onto unions of linear spaces (MRA), where we select the most representative coefficients in orthonormal wavelet bases, adapting to each case.

As a reminder, in the linear case, the notion of regularity of the function f is intimately related to the decay of the Fourier coefficients $|\hat{f}(\omega)|$. This regularity is global, meaning that even the slightest local discontinuity governs the decay of $|\hat{f}(\omega)|$, thus *primarily* degrading the low-dimensional approximation and *secondarily* "masking" the fact that the function may be very regular outside of this singular point. The question then arises: can we do better, and how? The answer is yes; we need to be able to localize the information of these discontinuities/transients. Note that it is not trivial to do this because these singularities/transients carry meaning (e.g., edge detection, note attack detection, etc.).

We have also seen that these linear/non-linear approaches provide entry points to understand single-layer neural networks with M neurons (Sec. 5.2). In particular, the linear approach allows us to understand the *Universal Approximation Theorem* by analogy with a Fourier series expansion and a change of basis that can be schematized by the "cosine to ReLU" formula. And we also understand that this theorem is futile because it is demonstrated that for regular functions, we have [55]

$$f \in L^2 \Rightarrow \lim_{M \to \infty} \|f - f_M\| = 0 \qquad \text{(Universality)} \qquad (232)$$

$$f \in H^\alpha \Rightarrow \|f - f_M\| = o(M^{-\alpha/d}) \qquad \text{(Curse)} \qquad (233)$$

In the non-linear framework, we have also seen A. Barron's approach, which involves adapting for each function f to obtain a sparse representation and taking only the most significant coefficients to define the low-dimensional approximation. This approach seems consistent to understand that, indeed, in the case of a neural network, it will be trained to answer a specific question: recognizing a cat image among images of dogs, coffee cups, etc., recognizing the sound of a piano among that of a violin, harp, etc. In particular, we use the ℓ^p ($p < 2$) norm, preferably $p = 1$ for sparsity. If we ensure that we control the Fourier coefficients of the function, then (Th. 10)

$$\sum_{n \in \mathbb{Z}^d} |\langle f(x), e_n(x) \rangle|^p \leq \infty \Rightarrow \|f - f_M\| = o(M^{-2/p+1}) \qquad \text{(Independence of } d\text{)} \qquad (234)$$

there is no longer a curse of dimensionality; the convergence is independent of the dimension d. However, again, this theorem is futile. Why? The reason is simple: the theorem

[55] H^α is a Sobolev space associated with the factor α, i.e., the "order of derivation."

tells us that everything goes well if the function is sparse in Fourier. However, images of cats, dogs, etc., or musical frames or voice spectra are not sparse at all in Fourier, and this approach does not explain at all the performance of deep neural networks.

Nevertheless, low-dimensional adaptive techniques are much more powerful than Fourier analyses for detecting transients. Thus, we have reviewed the implementation of multiresolutions and finding orthonormal wavelet bases (Sec. 7.1). We have revisited the RAP triangle by considering the extension of the notion of regularity in the sense of Sobolev. The key point is to capture localized singularities [56]. As in the Fourier case, we are led to calculate the correlations between the function and wavelets that are dilated/contracted and translated with zero mean ($\psi_{u,s}$, $\int \psi(u)du = 0$); these are the wavelet coefficients $W_f(u,s)$. The intensity of these coefficients indicates in the (u,s) plane both where (u_0) and at what scale (s_0) the function varies. We have seen that if we impose that the wavelet ψ has zero moments, it ignores polynomial parts of the signal, and the intensity of the coefficients W_f reflects deviations from this polynomial shape.

However, in order *primarily* to obtain low-dimensional approximations and *secondarily* to understand the link between regularity and the decay of coefficients W_f, we have set up representations that allow signal reconstruction by sampling in the (u,s) plane: $s = 2^j$ and $u = n2^j$ with $(j,n) \in \mathbb{Z}^2$. This leads to the orthonormal wavelet bases $\{\psi_{j,n}(u) = 2^{-j/2}\psi(2^{-j}u - n)\}_{(j,n)\in\mathbb{Z}^2}$, which allow a generalization of Shannon's sampling theorem (Th. 16). In the Fourier plane, the $\hat{\psi}_j(\omega) = \hat{\psi}(2^j\omega)$ define more or less dilated band-pass filters in which the spectrum $\hat{f}(\omega)$ is analyzed. The key point for the reconstruction of f to be possible is that the set of filters must cover the entire Fourier plane (*Littlewood-Paley condition*), namely $\sum_{j\in\mathbb{Z}} |\hat{\psi}_j(\omega)|^2 = 1$.

Historically, we knew the Haar and Shannon wavelets (Sec. 7.1.2), which are indeed orthonormal bases but are either discontinuous in real space for the former or an ideal band-pass for the latter, resulting in slow decay in $1/u$ in real space. The possibility of obtaining solutions with rapid and regular decay was long considered infeasible. However, we have seen that we can construct such orthonormal bases. Firstly, with a wavelet ψ that is both C^∞ and rapidly decreasing, as done by Y. Meyer (Sec. 6.3.3), then with wavelets constructed from nested sets V_j providing a linear approximation of the signal $P_{V_j}f$ at a scale of 2^j and whose orthonormal bases are derived by scale transformation (S.

[56] note that we are still restricting ourselves to cases where the number of singularities is not large

Mallat/Y. Meyer). From there, wavelets with *compact support* and possessing a certain number of *zero moments* have been constructed (I. Daubechies, Sec. 8.1). In this context, the wavelets ψ_j allow us to capture the details of the signal f at a scale of 2^j. These details complement the low-dimensional approximation $P_{V_j}f$, which is the essence of the decomposition:

$$P_{V_{j-1}f} = P_{V_j f} + P_{W_j f} \tag{235}$$

We then demonstrate that there exists a *scaling function* ϕ that allows us to obtain an orthonormal basis of V_j, namely $\{\phi_{j,n}(u) = 2^{-j/2}\phi(2^{-j}u - n)\}_{n\in\mathbb{Z}}$, and that there exists a *wavelet* ψ that gives an orthonormal basis of W_j, namely $\{\psi_{j,n}(u) = 2^{-j/2}\psi(2^{-j}u-n)\}_{n\in\mathbb{Z}}$. And finally, the union of all the bases $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$ forms an orthonormal basis for the space $L^2(\mathbb{R})$. The two functions ϕ and ψ are related by a relation seen in the course of theorem 18 (Sec. 7.4). The underlying key point is that the *scaling function* ϕ and the wavelet ψ are determined by the properties of two 2π-periodic filters $h(\omega)$ and $g(\omega)$, respectively:

$$\hat{\phi}(\omega) = \frac{1}{\sqrt{2}}\prod_{p=1}^{\infty}\hat{h}(2^{-p}\omega) \qquad \hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}(\omega/2)\hat{\phi}(\omega/2) \tag{236}$$

and $h(\omega)$ and $g(\omega)$ are related to each other by

$$\hat{g}(\omega) = e^{-i\omega}\hat{h}^*(\omega+\pi) \tag{237}$$

with $\hat{h}$ satisfying

$$\forall\omega\in[0,\pi/2],\ \hat{h}(\omega)\neq 0 \qquad |\hat{h}(\omega)|^2 + |\hat{h}(\omega+\pi)|^2 = 2 \tag{238}$$

Thus, the wavelet ψ is the result of a cascade of low-pass filters at different scales followed by a band-pass filter. From the wavelet basis of L^2, we can project any function f as follows:

$$f = \sum_{(j,n)\in\mathbb{Z}^2}\langle f,\psi_{j,n}\rangle\psi_{j,n} \tag{239}$$

and obtain the wavelet coefficients $W_f(j,n) = \langle f,\psi_{j,n}\rangle$. These wavelet coefficients are essentially zero (sparse representation) except at the locations of the singularities/transitions of the signal.

Now, what has been important from a practical point of view (at least) is the development of fast algorithms by S. Mallat for discrete wavelet transformations DWT and its inverse IDWT, with a complexity of $O(N)$, which is faster than the $O(N \log N)$ complexity of the FFT (Sec. 8.2). These algorithms are based solely on the properties of the Fourier coefficients, $h[n]$ and $g[n]$ $(n \in \mathbb{Z})$, of the filters $\hat{h}(\omega)$ and $\hat{g}(\omega)$. We perform a cascade of filtering and downsampling during a DWT (Fig. 47), and an inverse cascade in IDWT (Fig. 49). However, the structure of these cascades is derived from the nested structure of the sets V_j and their complements W_j. During the decomposition (DWT), the coefficients obtained via the h filter are low-frequency components of the signal, while the coefficients obtained via the g filter are high-frequency components.

This scheme has been extended (Y. Meyer) to 2D for image processing, but with a subtlety (Sec. 8.4). To obtain an orthonormal basis of $L^2(\mathbb{R}^2)$ from the scaling function ϕ and the wavelet ψ, we must now define 3 wavelets $\{\psi^k(u_1, u_2)\}_{k \leq 3}$. These three wavelets are obtained to cover regions of the Fourier space (Fig. 55) that are sensitive to transients: either horizontal, vertical, or having both types of variations. In dimension D, we need $2^D - 1$ wavelets. To complete the decomposition, we need to define $\phi(u_1, u_2)$ as a simple product of $\phi(u_1)\phi(u_2)$. Thus, we can implement a fast pyramidal algorithm (Fig. 54), where this time the high-frequency part has three components. As in 1D, most of the wavelet coefficients are zero, and only those that indicate a transient in their selection domain are significant.

Finally, we can relate the analysis of signal regularity to the behavior of wavelet coefficients. If we consider local Lipschitz-α/Hölder regularity, that is, for example, at u_0

$$|f(u) - f(u_0)| \leq C|u - u_0|^\alpha \tag{240}$$

depending on the value of α, we obtain different types of behaviors (Sec. 5.3.2). One way to view Lipschitz-α regularity is to quantify the amplitude of the increment with respect to $u = u_0$, that is:

$$|f(u_0(1 + s)) - f(u_0)| \leq C'|s|^\alpha \tag{241}$$

which results in a scaling of $|s|^\alpha$. Therefore, we relate the regularity property to the behavior of the signal during a dilation/contraction, which makes it quite intuitive that the wavelet transformation is well-suited to capture this type of regularity. Indeed, we have formulated the theorem (S. Jaffard, Th. 12) that can be written in the context of dyadic grids as follows:

$$\text{If } f \text{ Lipschitz}-\alpha \text{ at } u_0 \Rightarrow |\langle f, \psi_{j,n}\rangle| \leq C2^{(\alpha+1/2)j}(1+|n-2^{-j}u_0|^{\alpha}) \tag{242}$$

and, inversely, if for $\alpha' < \alpha$ we have

$$|\langle f, \psi_{j,n}\rangle| \leq C2^{(\alpha+1/2)j}(1+|n-2^{-j}u_0|^{\alpha'}) \tag{243}$$

then f is Lipschitz-α at u_0. Thus, we understand that the most important coefficients are localized at scale 2^j in the cone where the supports of the wavelets ψ_j have a non-zero overlap with the singularities of the signal. For example, we know that moving towards high frequencies means making s tend to 0, and thus j tends to $-\infty$, and the larger α is, the faster the coefficients decay. We then understand the wavelet decompositions such as the one shown in figure 46. By the way, if the function is bounded, which corresponds to $\alpha = 0$, the wavelet coefficients decay at least as $2^{j/2}$ (recall: going to high frequencies means $s \to 0$, which is equivalent to $j \to -\infty$). From these decompositions, we can conceive approximations of the signal f. First, we can keep only the low-frequency approximations (*linear approximation*), which correspond to projections onto the spaces V_j, i.e., $P_{V_j} f$ as illustrated in figure 51. In this case, we fix the size of the grid onto which we project the signal, which is similar to the strategy used when we want to compute a better basis, for example in PCA analysis. However, we observed the same type of problems as in Fourier, namely Gibbs oscillations at the locations of the singularities/transitions of the signal, because we only use the low frequencies. To do better, we resorted to the sparse description, which adapts (*non-linear approximation*) to the function f by requiring that only the wavelet coefficients greater than a threshold T be kept:

$$f_T = \sum_{|\langle f, \psi_{j,n}\rangle| > T} \langle f, \psi_{j,n}\rangle \psi_{j,n} \tag{244}$$

By doing so, we were able to verify that whether in 1D (Fig. 52) or in 2D (Fig. 56), we can reconstruct the original function/image with its fine details by keeping only about 10% of the total wavelet coefficients."

9.2 Quantitative Improvement in Nonlinear Approximation

Let's consider the following theorem [57]:

> **Theorem 19** *(Linear Framework)*
> *If $x \in C^\alpha[0,1]$ (Lipschitz-α), and if we keep only M "low-frequency" coefficients, then $\exists C$ such that:*
> $$\varepsilon_\ell^{1D} = \|x - x_M\|^2 \leq CM^{-2\alpha} \tag{245}$$

Proof 19.
Since we are in the linear framework, the approximation x_M is obtained by projecting x onto a linear space V_L, namely $P_{V_L}x$. We only need to adjust the size L to keep only M coefficients. Now,

$$P_{V_L}x = \sum_{n \leq M} \langle x, \phi_{L,n}\rangle \phi_{L,n} \tag{246}$$

with

$$\phi_{L,n}(u) = \frac{1}{\sqrt{2^L}}\phi(2^{-L}u - n) \tag{247}$$

As the support of x is $[0,1]$, then [58] $n \in [0, 2^{-L}]$. So, $M = 2^{-L}$.

Now, regarding the error, since we have the following decomposition of $L^2(\mathbb{R})$ (Sec. 7.4):

$$L^2(\mathbb{R}) = V_L \bigoplus_{j=-\infty}^{L} W_j$$

we can express the error of keeping only the projection onto V_L as:

$$\|x - x_M\|^2 = \sum_{j=-\infty}^{L} \sum_{n=0}^{2^{-j}} |\langle x, \psi_{j,n}\rangle|^2 \tag{248}$$

[57] Note: If in the previous section, the function notation was $f(u)$ to match S. Mallat's slides in the session, here we revert to the notation $x(u)$.

[58] Note: $L < 0$ because 2^L is the sampling rate on $[0,1]$, and we take $2^L \gg 1$.

Now, we know that the wavelet coefficients decrease as follows:

$$|\langle x, \psi_{j,n} \rangle| \leq C 2^{j(\alpha+1/2)} \tag{249}$$

So,

$$\|x - x_M\|^2 \leq C^2 \sum_{j=-\infty}^{L} \sum_{n=0}^{2^{-j}} 2^{(2\alpha+1)j} = C'(2^{2\alpha L}) = C' M^{-2\alpha} \tag{250}$$

■

This is the same result we obtained in Fourier analysis (Th. 6) for smooth functions. However, our interest lies in cases where the function is irregular.

It is interesting to study the 1D case because there is **no cost in encoding singularities**. Suppose the signal has Q singularities $u_1, \ldots, u_Q$ whose locations are unknown. The question is: how many wavelet coefficients will be affected by these singularities? Let's assume that x is C^α on each interval $]u_k, u_{k+1}[$. Inside these intervals, the wavelet will be "transparent", and we can apply the previous theorem. The coefficients will be large only around the discontinuities (i.e., in the "cone" discussed earlier). Now, if the wavelet has finite support, we can convince ourselves that for each scale 2^j, the number of translated wavelets whose support contains a singularity is constant, denoted as K. Thus, the total number of coefficients affected is: $QK \times N_s$, where N_s is the number of scales s that we retain. However, the latter is equal to $|L'|$ if we cut the decomposition at a scale L' with high frequencies. So, the number of coefficients affected by the singularities is:

$$M_1 = QK|L'|$$

What is the value of L'? If the signal is at least bounded on its support (including singularities), which corresponds to $\alpha = 0$, then the wavelet coefficients at scale 2^j decrease at least as $2^{j/2}$. But if we remove scales $j \in]-\infty, L']$, we make an error that can be quantified as follows:

$$\sum_{j=-\infty}^{L'} (QK) C^2 2^j = C^2 QK 2^{L'+1} = C' 2^{L'} \tag{251}$$

Now, we want this precision to be of the same order as in the linear case, i.e., $C' 2^{L'} \approx$

$M^{-2\alpha}$. This implies $|L'| \sim 2\alpha \log M \ll M$. So, the total number of coefficients is equal to those related to linear approximation and those related to the presence of singularities. Thus, the total number is roughly:

$$M_{tot} \sim M + C'' \log M$$

Therefore, by increasing the number of coefficients very slightly, the error in the presence of singularities is of the same order as the error without singularities.

In conclusion, we arrive at the following theorem:

> **Theorem 20** *(Nonlinear Framework)*
> *If $x \in C^\alpha$ on each interval $]t_k, t_{k+1}[$ $(k = 1, \dots, K)$, by keeping the M largest wavelet coefficients, there exists $C > 0$ such that:*
>
> $$\varepsilon_{n\ell}^{1D} = \|x - x_M\|^2 \leq CM^{-2\alpha} \tag{252}$$

So, when approximating a function with singularities in a linear framework, what governs the error decay is the smallest α in the interval $[0, 1]$. We practically end up with $\alpha \approx 0$ (bounded function) due to the step-like singularities, which would result in a decay of $1/M$ (unproven proposition). On the other hand, with a nonlinear approach, we are not limited by this slow decay, and α can be larger. Thus, we understand why a nonlinear approach in 1D gains significantly compared to a linear one.

However, **the phenomenon changes completely in higher dimensions**. If we revisit Theorem 19 and calculate the number of coefficients, we need to place ourselves on a grid. In 2D, we have $M = 2^{-2L}$. Also, in 2D, we know that the normalization of wavelets changes from $1/\sqrt{2^j}$ to $1/2^j$ to normalize the L^2 norm. Thus, the condition on wavelet coefficients is modified to:

$$|\langle x, \psi_{j,n} \rangle| \leq C 2^{j(\alpha+1)} \tag{253}$$

And if we look at the linear error, we must consider not 2^{-j} coefficients n but 2^{-2j} coefficients (n_1, n_2), so:

$$\varepsilon_\ell^{2D} = \|x - x_M\|^2 \leq C^2 \sum_{j=-\infty}^{L} 2^{-2j} 2^{2j(\alpha+1)} \leq C' 2^{2\alpha L} = C' M^{-\alpha} \tag{254}$$

Generalizing to dimension q is straightforward, and we see that:

$$\varepsilon_\ell^{2D} \leq C' M^{-2\alpha/q} \qquad \text{(no singularities)} \tag{255}$$

Now, if we consider a function that is only C^α piecewise, i.e., with singularities, there is a change because even nonlinear approximation will be limited. Let's take the example of the square image (Fig. 54) and more precisely a horizontal edge of the white square. We'll trace the reasoning in 1D, identifying the differences. The number of wavelets that will be sensitive to the edge is roughly:

$$M_j = \ell/2^j \times K \tag{256}$$

where ℓ is the size of the edge, 2^j is the number of translations along the edge of a wavelet at this scale, and K is the number (constant) of wavelets involved in the vertical direction (1D reasoning). Now, the function has a bounded step-like structure, so $\alpha = 0$. This imposes a constraint on the wavelet coefficients (Eq. 253):

$$|\langle x, \psi_{j,n} \rangle| \leq C 2^j \tag{257}$$

Hence, the linear error in the presence of this edge is roughly:

$$\varepsilon_{n\ell}^{2D} \leq \sum_{j=-\infty}^{L'} M_j C^2 2^{2j} = (\ell K C^2) 2^{L'} \tag{258}$$

However, $|L'|$ is still of the order $\log M$, but $M = 2^{-2L}$, so $|L'| \sim 2|L|$. Thus, we obtain:

$$\varepsilon_{n\ell}^{2D} \leq C' M^{-1} \qquad \text{(with singularities)} \tag{259}$$

This implies that even if the image is C^∞ (e.g., inside/outside the white square), **the nonlinear error is dominated by the discontinuities**. As we increase the dimension q, the surface of the discontinuities increases, and so we need to increase the number of coefficients to achieve good approximation.

9.3 Compression

Let's now consider the perspective of the amount of information we need to retain in order to reconstruct an image effectively. From a naive point of view, we might think that we can encode each pixel of the image either with, for example, 8 bits (256 levels of grayscale) or 1 bit (B & W). Transitioning from 8 bits to 1 bit obviously reduces the size by a factor of 8, but it impacts the quality, and we lose a lot of information (Fig. 57).

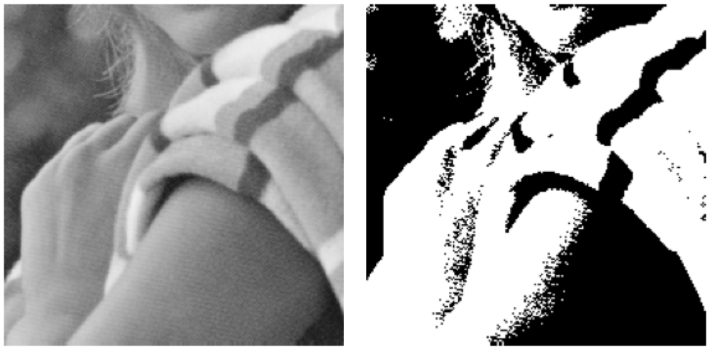

FIGURE 57 – On the left, an image where each pixel value is encoded with 8 bits (1 byte), and on the right, the same image with pixel values encoded using 1 bit.

However, we can do better. The idea is as follows: we work in the orthonormal wavelet basis, and we compress the wavelet coefficients (function $Q(x)$) to a few bits:

$$\tilde{x} = \sum_{j,n} Q(\langle x, \psi_{j,n} \rangle) \psi_{j,n} \tag{260}$$

Now, if we look at the distribution of wavelet coefficient values (details only) in Figure 58, we see that the distributions are indeed peaked at 0 (illustrating sparsity), but the spread varies depending on the complexity of the textures. Therefore, it's natural to work in the wavelet basis that provides sparse representations with many zeros. Initially, we only encode the non-zero coefficients. To do this, we use a fixed-step quantizer with the following definition: if the value x takes values in $[-a, a]$, and this interval is divided

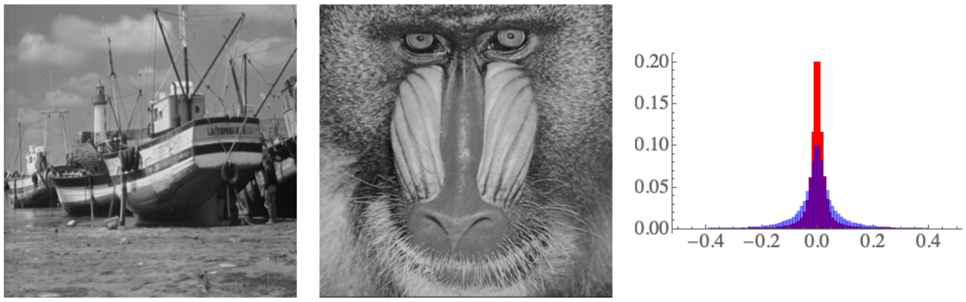

Figure 58 – On the left and in the middle, two images. On the right, in red, the normalized histogram of wavelet coefficients (details only) for the boat image, and in blue for the monkey image, which exhibits more textures.

into $n = 2p + 1$ boxes, then:

$$Q_\Delta(x) = k\Delta \qquad \Delta = \frac{2a}{n} \qquad k = \lfloor x/\Delta + 1/2 \rfloor \in [\![-p, p]\!] \tag{261}$$

In other words, we assign the central value of the k-th box to x. Therefore, a coefficient whose value lies in the interval $[-\Delta/2, \Delta/2]$ is assigned a value of 0. This is particularly interesting when considering the histogram of coefficient values (Fig. 58). Thus, we construct a binary map $b[j, n]$ such that:

$$b[j, n] = \begin{cases} 0 & \text{if } Q_\Delta(\langle x, \psi_{j,n} \rangle) = 0 \\ 1 & \text{otherwise} \end{cases} \tag{262}$$

An example is shown in Figure 59. If we now assign a value to the non-zero coefficients using the simple quantizer above, we obtain reconstructed images for the parameters shown in Figure 59. To appreciate the difference between the original image and two reconstructions with different values of p, the first being the one in Figures 59 and 60, and the other four times larger (with bins four times smaller), we show zoomed-in sections of the reconstructed images and the original images in Figure 61.

The JPEG2000 standard employs a similar approach with refinements that would be too lengthy to mention here (see S. Mallat's book). However, the key message is that with a compression factor of 40 (meaning we go from 8 bits/pixel to 0.2 bit/pixel), we

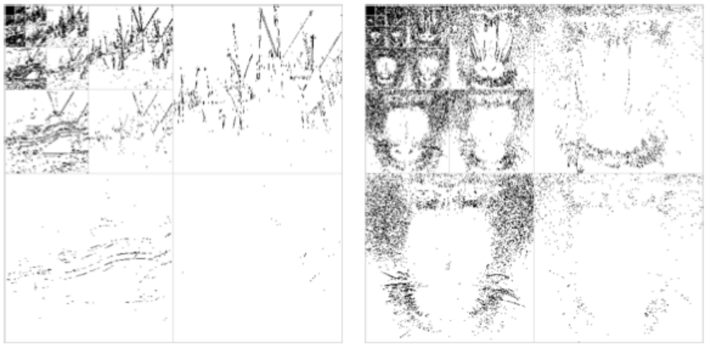

FIGURE 59 – For the two images in Figure 58, we applied binarization (Eq. 262) with the parameters $a = 5$ and $p = 15$ from the quantizer Eq. 261 (note that the distributions of detail coefficients have fairly long tails). Coefficients resulting in a value of 0 appear as white in the images. Wavelet decomposition is performed with "Db2".

can faithfully reconstruct the image. Of course, if we reduce the number of bits/pixel too much, we start to lose details, as observed in Figure 60. Typically, an image can be compressed by a factor of 50-100 without introducing significant artifacts. For higher compression ratios, these techniques do not suffice.

That being said, how do these observations relate to nonlinear approximation? The approximation error of the image x by the coding $\tilde{x}$ (Eq. 260) can be described in the orthonormal wavelet basis. If we use the simple quantizer Q_Δ, it follows:

$$\begin{aligned} \|x - \tilde{x}\|^2 &= \sum_{j,n} |\langle x, \psi_{j,n}\rangle - Q_\Delta(\langle x, \psi_{j,n}\rangle)|^2 \\ &= \sum_{|\langle x, \psi_{j,n}\rangle)| \leq \Delta/2} |\langle x, \psi_{j,n}\rangle|^2 + \sum_{|\langle x, \psi_{j,n}\rangle)| > \Delta/2} \left(\frac{\Delta}{2}\right)^2 \end{aligned} \tag{263}$$

The first sum is the error when we keep only the M largest coefficients. Moreover, the number of coefficients not set to 0 by quantization (the second part of the sum) is exactly M due to the same argument. Therefore, the distortion D due to coding is bounded by:

$$\varepsilon_{n\ell}(M) \leq D = \|x - \tilde{x}\|^2 \leq \varepsilon_{n\ell}(M) + \frac{M\Delta^2}{4} \tag{264}$$

Hence, this distortion D has two components: a nonlinear component due to the coeffi-

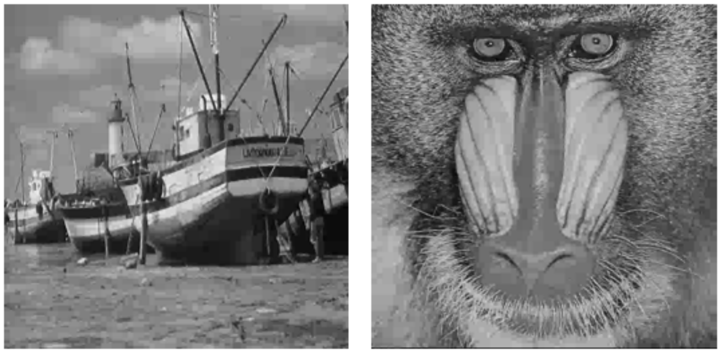

Figure 60 – Reconstructed images obtained from quantizing wavelet coefficients with the parameters from Figure 59. The effect of quantization is visible.

cients not retained by the thresholding induced by the bin "0" width, and a component due to the quantization of the coefficients retained by the nonlinear approximation.

If we assume that the wavelet coefficients are sparse (as suggested by the histograms in Figure 58), which can be translated into a bounded ℓ^p norm (Sec. 5.1.2), for example:

$$\sum_{j,n} |\langle x, \psi_{j,n}\rangle)|^p \leq C_p^p \tag{265}$$

then the k-th wavelet coefficient (Th. 9) satisfies:

$$|\langle x, \psi_k\rangle)| \leq C_p^p k^{-1/p} \tag{266}$$

The number M corresponds to the number of coefficients whose absolute value is greater than $\Delta/2$. Thus:

$$\Delta/2 = C_p^p M^{-1/p} \tag{267}$$

and according to the same theorem:

$$\varepsilon_{n\ell}(M) \leq \frac{C_p^{2p}}{2/p - 1} M^{1-2/p} \tag{268}$$

Therefore:

$$D = \|x - \tilde{x}\|^2 \leq \frac{C_p^{2p}}{2 - p/2} M^{1-2/p} \tag{269}$$

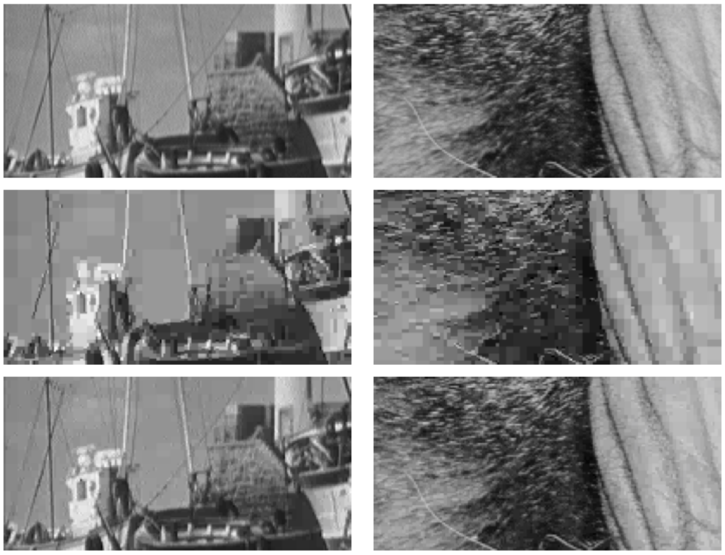

Figure 61 – Zoom: at the top, the original images; in the middle, images from Figure 60; and at the bottom, images reconstructed with quantization having four times more bins.

Hence, the quantization error and the nonlinear error are of the same order, implying that $D = O(M^{1-2/p})$. This tells us that **the coding error essentially depends on the low-dimensional nonlinear approximation**. This is because the number of coefficients set to 0, contributing to the nonlinear error, is very large due to sparsity. The number of bits required to encode the information has two components, both proportional to M, one arising from the localization of zero coefficients and the other from encoding the amplitudes of the non-zero coefficients. The crucial point to remember is that **these coding algorithms require a sparse representation**.

*Next year, we will delve into **Information Theory*** to understand how neural networks operate. In high dimensions, we have a valuable asset: the Central Limit Theorem, which tells us that when summing independent variables, we converge towards the mean. This phenomenon indicates that information is concentrated in certain parts of space, and the number of bits required to encode it is defined by **entropy**.