united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

SMR1237/3

School on Mathematical Problems in Image Processing

(4 - 22 September 2000)

Wavelets and functions with bounded variation from image processing to pure mathematics

Y. Meyer
C.M.L.A.
Ecole Normale Supérieure de Cachan
61, avenue du President Wilson
F-94235 Cachan cedex
France

These are preliminary lecture notes, intended only for distribution to participants

Wavelets and functions with bounded variation from image processing to pure mathematics

Yves Meyer

One goal of these lectures is to provide you with some information about what is happening in the still image compression standard which is developped under the name of JPEG-2000. A second goal consists in explaining how this technological challenge is motivating some advances in pure mathematics. This will eventually lead to some remarkable improvements on the Gagliardo-Nirenberg inequalities.

Explaining the performances of JPEG-2000 requires a model for still images. Among several models, the one on which this discussion is based was proposed by Stan Osher and Leonid Rudin. In this model, the simplified image is assumed to be a function with bounded variation. Then the efficiency of wavelet based algorithms will be related to the remarkable properties of wavelet expansions of functions with bounded variations (BV).

Once the fundamental properties of wavelet coefficients of BV-functions are better understood, this will yield some new Gagliardo-Nirenberg estimates. As it was pointed out by Albert Cohen and Ron DeVore, a similar strategy can be used for proving the "averaging lemma" of P.L. Lions and Ron DiPerna.

This lecture is organized as follows. In section 1, a few experiments are illustrating the efficiency of wavelets based algorithms in still image compression technology. In section 2, some terminology used in image compression will be clarified. In section 3, $u + v$ models for still images are introduced. Best-basis algorithms are introduced in section 4. Section 5 is devoted to the already obsolete JPEG standard. This standard is often justified as being a particular example of a general methodology, named principal component analysis or Karhunen-Loève expansion, a concept which is analyzed in section 6. In section 7, the limitations of such a Karhunen-Loève approach to

compression will be illustrated with a counter-example. In section 8, we will return to the $u + v$ models for still images on which this work is based. In section 9, some well known properties of functions with bounded variations (BV) will be listed for the reader convenience. In section 10, this space BV is used for modeling still images. A survey of wavelet analysis will be provided in section 11. Then in section 12, quantization issues will be addressed. Fourier series expansions of BV functions will be compared to their wavelet expansions in section 13. Finally in section 14, the preceding tools will be applied to improving Gagliardo-Nirenberg estimates.

This talk is based on a joint work with Albert Cohen and Frederic Oru.

1 Wavelets and still image compression

Let us begin with some examples of technological applications of wavelets. The first example is extracted from the web page of the Pegasus company (http://www.jpg.com). It reads the following:

> Pegasus Imaging Corporation has partnered with Fast Mathematical Algorithms & Hardware Corporation and Digital Diagnostic Corporation to develop new wavelet compression technologies designed for applications including medical imaging, fingerprint compression, video compression, radar imaging, satellite imaging and color imaging.
>
> Pegasus provides wavelet compression technology for both medical and non-medical application. Pegasus' wavelet implementation has received FDA market clearance for medical devices.
>
> This software is the only FDA-approved lossy compression software for image processing. Recent clinical studies have shown that the algorithm is comprehensively superior to other similar compression methods. It is licensed to multiple teleradiology developers and medical clinics including the Dutch software vendor Applicare Medical Imaging and the UK telecom giant British Telecom.

The second advertisement comes from a company named "Analog Devices". It reads:

> Wavelet compression technology is the choice for video capture and editing. The ADV601 video compression IC is based on a mathematical breakthrough known as wavelet theory...This compression technology has many advantages over other schemes. Common discrete schemes, like JPEG and MPEG, must break an image into rectangular sub-blocks in order to compress it... Natural images highly compressed with DCT schemes take on unnatural blocky artifacts...Wavelet filtering yields a very robust digital representation of a picture, which maintains its natural look even under fairly extreme compression. In sum ADV601 provides breakthrough image compression technology in a single affordable integrated circuit.

A third success story tells us about the FBI and fingerprints. It says:

> The new mathematical field of wavelet transforms has achieved a major success, specifically, the Federal Bureau of Investigation's decision to adopt a wavelet-based image coding algorithm as the national standard for digitized finger-print records... "

The interested reader is referred to a paper by Christopher Brislawn in the Notices of the AMS, November 1995, Vol 42, Number 11, pages 1278-1283 or to the remarkable web site of Christopher Brislawn [2].

Our next advertisement for wavelet-based image compression is coming from the celebrated Sarnoff Research Center. It reads:

> A simple, yet remarkably effective, image compression algorithm has been developed, which provides the capability to generate the bits in the bit stream in order of importance, yielding fully hierarchical image compression suitable for embedded coding or progressive transmission...

Finally the last example will concern the upcoming JPEG-2000 still image compression standard. While the JPEG committee is still actively working, il is very likely that the JPEG-2000 standard will be based on a combination of wavelet expansion (the choice of the filter is not fixed, and could include biorthogonal filters such as the 9/7, as well as 2-10 integer filters) and trellis

coding quantization. Applications range from Medical imagery, client/server application for the world wide web, to electronic photography and photo and art digital libraries.

These examples are showing that still image compression is a rapidly developing technology with far reaching applications.

2 A first glance to still image compression

We now turn to some mathematical models for image compression. We start with the superficial approach that an analog image on a domain D can be viewed as a function $f(x_1, x_2) = f(x)$ belonging to the Hilbert space $H = L^2(D)$. The energy of such an image is, by definition, $\int_D |f(x)|^2 dx$. It is obvious that an arbitrary such function $f(x)$ in H is far from being a natural image or something looking similar but this hot issue will be later clarified. Indeed our main problem will be to try to understand how an image differs from an arbitrary L^2 function.

For sampling this analog image into a digital image we need to fix a grid which is defined as being $N^{-1}Z \times N^{-1}Z$ for some large N. We then speak of a fine grid. Generally $N = 2^j$ and the corresponding grid is denoted by Γ_j. These Γ_j are embedded since Γ_j is contained in Γ_{j+1}. A digital image f_j is now a matrix indexed by points in Γ_j and, at this time of the discussion, we do not want to bother the reader with a precise definition of the mapping P_j which maps f to f_j. This mapping is certainly not a simply minded restriction from $L^2(D)$ onto $l^2(\Gamma_j)$ since this restriction does not make any sense. A smoothing is needed before sampling. Indeed these two operations are combined together inside a pyramidal algorithm. We then can move one step further and follow A. Rosenfeld who defines sampling by computing the coefficients of the image in some orthonormal basis Z. Indeed one can construct an analyzing wavelet ψ and a scaling function φ such that the effect of the pyramidal algorithm amounts to computing the coefficients of the expansion of the analogue image in the orthonormal basis generated by this wavelet and this scaling function. This point will be later clarified when wavelets will be defined (section 11)..

If D is the unit square, this digital image $f_j \in l^2(\Gamma_j)$ is now a huge matrix $c_{(k,l)} = f_j(x_1, x_2)$ where $x_1 = k2^{-j}, x_2 = l2^{-j}$, k and l ranging from

0 to $2^j - 1$. These entries $c_{k,l}$ are called pixels and each $c_{k,l}$ measures the gray level of the given image at $(k2^{-j}, l2^{-j})$. Here we are talking about a white and black image and a colour image has a similar definition with the difference that $c_{k,l}$ is now vector valued. Our digital image can be viewed as a vector inside a 4^j dimensional vector space. The gray levels $c_{k,l}$ are finally quantized with a eight bit precision which provides 256 gray levels. This discrete representation of an image needs to be compressed in order to be efficiently stored or transmitted.

Compressing still images means taking advantage of the local correlations between neighboring pixels in order to find some lower dimensional approximations. For instance, if a given pixel is bright and red, its neighboring pixels of are more likely to be bright red than dark blue. That is not to say that abrupt changes do not happen but they are mainly occuring on curves or one-dimensional subsets. These curves are often the edges of the objects to be detected in the image and are therefore providing some important information.

Compression might also based on some other geometrical properties of images. One may argue that an image represents objects which have a specific geometrical organization in the three dimensional space. The edges we find in an image tell us something about this organization. For example occlusions in the 3-D scene correspond to T-junctions for the corresponding edges. An other viewpoint was advocated by Benoit Mandelbrot who proposed self-similar stochastic models for image processing. These models led to beautiful simulations of natural landscapes.

All scientists working on image processing agree on the possibility of compressing images but they immediately diverge when one is asking about the precise mathematical description of the models for natural images on which the compression algorithms crucially rely.

A large group of scientists was following a method which was successful in mathematical physics. One team discovered the fundamental equation (PDE) which governs image processing. Indeed this equation (mean curvature motion) tells you what deformations or evolutions of an image are consistent with contrast changes. An other group proposed some plausible axioms (or laws) which are aimed to provide good models for natural images together with mathematical formulations of the specific and concrete tasks to be performed on some given collections of images. We will be more specific about these two approaches. Then many tasks in image processing can be

reformulated as problems in mathematics. This methodology uses tools and strategies which are quite familiar in other parts of applied mathematics and heavily relies on the resources of numerical analysis. Important examples are the Osher-Rudin model [23] on which this talk relies (this model will be presented in section 10) and the variational formulation of image segmentation which was proposed by David Mumford and Jayant Shah [21]. J.M.Morel and his school went one step further. They completely described and studied all nonlinear evolutions which are aimed to provide accurate sketches of a given image. These evolution laws are consistent with some specific invariance properties which are expected in image processing (contrast invariance etc) [18].

An other group of scientists is advocating for a more experimental study. They claim that stochastic models for collections of natural images should be learnt from large data sets. For example, a model used for the faces of male graduate students at MIT should differ from the one used for the faces of female students at NYU. In other words each collection of natural images should be given a separate stochastic model [20]. In this approach, simple and elegant fundamental laws simply do not exist and image processing tasks are performed by statistical methods. However these statistical methods are mathematically founded and this empirical approach to image processing will eventually become a piece of science when the statistical modeling of collections of natural images will be fully available.

Since large data sets of natural images are already at our disposal, compression algorithms which are based on a mathematically oriented axiomatization of image processing can be experimented and tested. These checkings are playing a crucial role since the choice of the axioms or fundamental laws suffer from some arbitrary. Moreover trained experts are needed for such checkings since there are no scientific criteria to measure the quality of a reconstructed image. The human vision system seems the best judge.

These experimental studies have been conducted in the five above mentioned examples. The experts acknowledged the improvements obtained by wavelet methods. The competing algorithm is the already obsolote JPEG standard. This celebrated JPEG algorithm will be visited in section 5. The difference between the old JPEG with wavelet methods will amount to comparing Fourier analysis to wavelet analysis.

3 A first visit to $u+v$ models for still images

Why do wavelet algorithms perform better than Fourier methods in image compression? One answer to this problem relies on an axiomatic model proposed by Osher and Rudin (among others). This model is named a $u+v$ model. This model is introduced in this section and will be revisited in section 8 and 10.

In a $u+v$ model, images are assumed to be a sum of two components $u(x)$ and $v(x)$. The first component $u(x)$ is modeling the objects or features which are present in the given image while the $v(x)$ term is responsible for the texture and the noise. But the textures are often limited by the contours of the objects and $u(x)$ and $v(x)$ should be coupled by some geometrical constraints. These constraints are absent from most of the $u(x)+v(x)$ models.

A main challenge in image processing consists in finding some relevant objects or features (some tanks, for example) inside a noisy image. In the $u+v$ models, this amounts to isolate the $u(x)$ component from the $v(x)$ component in a sum $f(x)=u(x)+v(x)$. Extracting $u(x)$ from $u(x)+v(x)$ can be compared to obtaining a drawing by Ingres from a painting by Renoir or Monet. This $u(x)$ component will provide a sketchy approximation to the given image. In the following discussion we will view the $u(x)$ component as a good approximation to the given image since $u(x)$ is expected to retain the most important features which are present in $f(x)$. Compression algorithms will be judged on their ability to preserve the $u(x)$ component of an image which is still waiting for a precise definition.

Precise stochastic models are not yet available for describing this $u(x)$ component and many scientists decided to switch to a functional space model. In a functional space model, this u component si assumed to belong to a ball of a functional Banach space B. Anticipating a discussion to be continued in section 10, let us now give a few examples of some Banach spaces B which are used. In the Osher-Rudin model [23], B is the space BV of functions with bounded variations. In the DeVore-Lucier model [8], B is a slightly smaller Besov space which admits a trivial wavelet characterization. Finally in the Mumford-Shah model [21], u is indeed a pair (u,K) where K is a collection of curves or more generally a one-dimensional set and u, once restricted to the complement of K, belongs to the Sobolev space H^1.

Let us now say a few words about the v component of an $u+v$ model. This v component is not structured and will be modeled by an arbitrary

function in $L^2(D)$. In all the $u+v$ models, the Banach space B which is used for modeling u is contained in $L^2(D)$ and the reader may wonder what is then the difference between the u and the v component. The answer is the following: the v component should have a small L^2 norm: $\|v\|_2 \leq \epsilon$ for some small ϵ. A second and more accurate answer is that v can both include a textured component with a small energy and an additive noise.

Let us now relate these $u+v$ models to quantization.

Quantization is a crucial segment of a compression scheme. In a compression scheme the first step maps a given image into a string of coefficients. These coefficients are real numbers and these numbers are replaced by some digital approximations depending on the computer precision. Once quantization is performed, the image can be transmitted and the reconstructed image will be affected by the quantization. The error which affects the reconstructed image heavily depends on the orthonormal basis Z which is used. We would like this quantization error to be less damageable to the $u(x)$ component than to the $v(x)$ one. In the ideal case, the $u(x)$ component should remain untouched while the $v(x)$ might disappear. Then quantization would de-noise the given image. If it is the case, then the function which is reconstructed from the quantized coefficients yields an approximation to the given image which keeps track of the main features (edges...). All these features were assumed to be present in the $u(x)$ component of the image. However the texture will unfortunately be treated as being noisy by this scheme and wiped out. To conclude, an image is always distorted after quantization and transmission and we would like to minimize this distortion by a best-basis search. This issue will be soon clarified.

Thresholding is closely related to quantization and raises fascinating mathematical problems. We will accept the working hypothesis that what is happening to an image after thresholding gives a good indication on what happens after quantization. Let us define thresholding. A given signal or image $f(x)$ is decomposed in some orthonormal basis Z of Hilbert space H. This reads $f(x) = \sum_0^\infty c_n e_n(x)$. The coefficients c_n that appear in this expansion are sorted out. All coefficients c_n such that $|c_n| < \epsilon$ where ϵ is a given threshold are viewed as being insignifiant and replaced by 0. In short we define $q_\epsilon(x) = x$ if $|x| > \epsilon$ and $q_\epsilon(x) = 0$ otherwise and the thresholding operator

Q_ϵ is defined by

$$Q_\epsilon(f) = \sum_0^\infty q_\epsilon(c_n) e_n(x) \tag{3.1}$$

Then the resulting error measured in the H norm is $R_\epsilon = \|Q_\epsilon(f) - f\|_H$. Computing other errors $\|Q_\epsilon(f) - f\|_B$ for more general Banach spaces B is a fascinating problem which will be later addressed (Theorem 4). Returning to the Hilbert space case, the error R_ϵ is easily estimated if the coefficients $|c_n|$ are sorted out (or rearranged) as a non-increasing sequence c_n^*. We then obtain

$$R_\epsilon = (\sum_{n>N} (c_n^*)^2)^{1/2} \tag{3.2}$$

where N is defined as the lowest index n such that $c_n^* < \epsilon$. The error R_ϵ depends on the threshold ϵ and on the decay of c_n^*. This decay depends on the signal and also on the specific orthonormal basis Z which is being used.

Quantization and thresholding are related issues where a key role is played by the orthonormal basis Z in which we expand the given signal. This leads to a best basis search as described in the following section.

4 Best-basis algorithms in signal processing

David Marr was insisting on the role played by the choice of a particular representation for achieving some signal processing tasks. He wrote in his famous book, "Vision":

> A representation is a formal system for making explicit certain entities or types of information, together with a specification of how the sytem does this. And I shall call the result of using a representation to describe a given entity a description of the entity in that representation. For example, the Arabic, Roman and binary numerical systems are all formal systems for representing numbers. The Arabic representation consists in a string of symbols drawn from the set 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and the rule for constructing the description of a particular integer n is that one decomposes n into a sum of multiple of powers of 10... A musical score provides a way of representing a symphony; the alphabet allows the construction of a written representation

> of words... A representation, therefore is not a foreign idea at all we all use representations all the time. However, the notion that one can capture some aspects of reality by making a description of it using a symbol and that to do so can be useful seems to me a fascinating and powerful idea...

Many orthonormal bases are currently used in signal processing. Sometimes the choice of a particular basis is not seriously justified. A Fast Fourier Transform (FFT) is cheap and available and one is tempted to use it without questioning. However many scientists took an opposite view and argued that one should relate the basis (or the representation in David Marr's language) to the tasks to be performed. If this task is compression, a systematic search for an optimal basis led to the celebrated Karhunen-Loève algorithm which will be analyzed in section 6 and 7.

Let us indicate how a best basis search for quantization and thresholding can be further conducted in the framework of a $u + v$ model.

If the $u+v$ model is accepted, we can write $f(x) = u(x)+v(x)$ where $u(x)$ represents the important features of the given signal or image while $v(x)$ represents the noise or some components of the signal which are less interesting. We do not want the $u(x)$ component to be damaged by a quantization or a thresholding. This amounts to finding a "better perspective" from which $u(x)$ and $v(x)$ would be clearly and neatly separated. This new perspective or view point will be provided by a new orthonormal basis $e_n(x), n \in N$, in which the coefficients of $u(x)$ should be clearly distinguishable from the coefficients of $v(x)$. If $v(x)$ is a white noise, then its coefficients in any orthonormal basis will be i.i.d. $N(0, \sigma)$ where σ is the noise level. In other words the coefficients v_n of $v(x)$ are extremely flat and in contrast we would like the coefficients u_n of $u(x)$ to plunge as fast as possible. Since the l^2 norm of these coefficients u_n does not depend on the orthonormal basis $e_n(x)$ which is used, the fact that u_n might plunge faster in a particular basis implies that they also peak higher. Using a metaphor which is familiar to climbers, the sequence u_n should resemble the high peaks which still receive the sunset light while the sequence v_n is like the valleys which are already buried in the darkness of the coming night.

Returning to signal or image prossessing, let us assume that an orthonormal basis $e_n(x)$ can be found such that the sorted coefficients of $u(x)$ have a fast decay. Then by thresholding the coefficients c_n of $f(x)$ in this basis,

we will retain much of the energy of the $u(x)$ component and wipe out $v(x)$. Similarly the expansion of u will be compressed with a few terms in this particular basis.

Faced to these challenges, wavelet analysis will perform much better than Fourier analysis under the following two conditions. The first condition says that an $u + v$ model based on a specific Banach space B is adapted to our problem and the second condition tells us that the wavelet coefficients of functions in B should plunge much faster than the corresponding Fourier coefficients. For instance the sorted wavelet coefficients of a function with bounded variation decay as $1/n$ (theorem 7) while the sorted Fourier coefficients of functions with bounded variation may decay as badly as $n^{-1/2}$ if we ignore logarithmic factors. If the Osher-Rudin model for still images is accepted, theorem 7 will explain the performances of wavelets in image compression.

5 The old JPEG

The old JPEG algorithm is based on the DCT (discrete cosine transform). The DCT is an alternative to the usual Fourier series expansion. We start with a fonction $f(x)$ in $L^2(R)$ that we want to analyze over the interval $[0, \pi]$. We then restrict the given function to this interval and get a function $g(x)$ in $L^2(0, \pi)$. We do not want to extend this new function as a π-periodic function of x, since this would create artificial discontinuities at the end points of our interval. Instead we extend $g(x)$ into a 2π-periodic function $h(x)$ by imposing the mirror conditions $h(x) = h(-x)$, $h(x) = g(x)$ over $[0, pi)$ and finally $h(x) = h(2\pi - x)$. Then this $h(x)$ can be written as a Fourier series. This is equivalent to directly starting with the orthonormal basis of $L^2(0, \pi)$ given by $e_0(x) = \sqrt{1/\pi}$ and $e_n(x) = \sqrt{2/\pi}\cos(nx), n \geq 1$. The discrete cosine transform (DCT) is nothing else but the discrete version of the preceding orthonormal basis. There are some variants on this DCT algorithm and one consists in extending $g(x)$ into a 4π-periodic function $h(x)$ which satifies $h(x) = h(-x), h(x) = g(x)$ over $[0, \pi]$ and finally $h(x) = -h(2\pi - x)$. It amounts to using the orthonormal basis of $L^2(0, \pi)$ given by $f_n(x) = \sqrt{2/\pi}\cos((n + 1/2)x)$.

The old JPEG consists in dividing the sampled image into 8×8 blocks. For a 512×512 image there are 4096 such blocks. Inside each block (defined

by $1 \leq x, y \leq 8$) the two-dimensional DCT is using the orthonormal basis given by the vectors

$$4^{-1} \cos(\pi(k+1/2)x) \cos(\pi(l+1/2)y), \qquad 1 \leq x, y, k, l \leq 8 \tag{5.1}$$

Finally the 64 resulting coefficients are quantized in each block. It means that the real numbers which arise as the coefficients in the DCT expansion are replaced by the nearest approximations which are compatible with the machine precision and the bit budget which we are allowed to spend. From this discussion we already observe that it is not reasonable to impose a compression factor larger than 100 to the old JPEG. In such a case all the coefficients in a given 8×8 block might disappear!

The old JPEG can be interpreted as a windowed Fourier analysis. It means that JPEG is computing a local frequency. If the given image is smooth on some 8×8 block, then most of the (local) Fourier coefficient will vanish. This is the only place where some compression might happen. In contrast a function with a sharp discontinuity along a curve is poorly compressed in such a Fourier basis.

Using DCT for still image compression is a decision which is often motivated by a more general paradigm concerning Karhunen-Loève (KL) bases. These KL bases are aimed to provide an optimal compression. This leads us to discuss the strength and the weaknesses of Karhunen-Loève expansions in the next section.

6 Karhunen-Loève expansions

The Kahrunen-Loève representation algorithm is based upon a mean square error approach to compression. We are given a collection $X_1, ..., X_n$ of vectors belonging to R^N (N is huge). For the sake of simplicity it will be assumed that $X_1 + ... + X_n = 0$. We want to compress a generic vector X_j belonging to this collection. By "generic" we mean that the index j is chosen at random in the probability space $\Omega = \{1, ..., n\}$ where the same probability $1/n$ is given to every j. We want to reduce the large dimensionality N to a much smaller one q. In a linear compression scheme it means that the compressed vectors $Y_1, ..., Y_n$ should belong to a q-dimensional vector space V_q. All errors are measured with the usual energy given by the euclidean structure of R^N

which implies that Y_j is the orthogonal projection of X_j onto V_q. Our goal is to minimize the mean square error which results from replacing X_j by Y_j when $1 \leq j \leq n$. As it was said before we consider that all X_j's , $1 \leq j \leq n$, have the same probability. This leads for a given q to finding the q-dimensional vector space V_q which offers an optimal fit with the cloud of vectors $X_1, ..., X_n$. In other words $\sum_{j=1}^{n} \|X_j - Y_j\|^2$ should be minimal. It is easily proved that these vector spaces V_q are embedded: $V_0 = \{0\}$ is simply the arithmetical mean (center of gravity) of our vectors $X_1, ..., X_n, V_1$ is the inertia axis and so on until we reach $V_N = R^N$.

The Karhunen-Loève basis then consists in the new orthonormal basis $e_1, ..., e_N$ of R^N defined by the property that for each q, $e_1, ..., e_q$ is a basis of V_q. A Karhunen-Loève expansion is the most efficient way to compress a collection of signals and images by a linear algorithm in which we want to minimize a mean square error. It should be stressed that one cannot compute a Karhunen-Loève basis if we are not given a stochastic model for the collection of signals or images we want to compress. Here this stochastic model consisted in agreeing that all j's were given the same probability.

Returning to image processing, some authors used the working hypothesis that most collections of natural images are translation invariant. This is obviously false in many cases. An obvious counter-example is given by the faces of human beings. Indeed we never accept large translations on such pictures but rather demand that the faces be correctly centered. If we accept this translation invariance, it implies that the trigonometric system is indeed the Karhunen-Loève basis. A Fourier expansion provides the best average compression for such translation invariant families of images. This is the argument which is used in favor of JPEG but it does not tell us why we should first segment our images into 8×8 blocks before using a DCT.

Before questioning the Karhunen-Loève algorithm, let us state that this algorithm can obviously be generalized to stochastic processes. We assume that $X(t, \omega)$, $a \leq t \leq b, \omega \in \Omega$, belongs to $L^2([a, b])$ where we are given an a priori probability law $d\omega$ on Ω. Then the Karhunen-Loève expansion of $X(t, \omega)$ reads

$$X(t, \omega) = \sum_{0}^{\infty} e_n(t) r_n(\omega) \tag{6.1}$$

where $e_n(t)$ is an orthonormal sequence in $L^2[a, b]$ while the sequence $r_n(\omega)$ is orthogonal in $L^2(\Omega, d\omega)$. Then the partial sums $\sum_{n=1}^{q} e_n(t) r_n(\omega), q \geq 1$

provide the best linear approximations in $L^2([a,b] \times \Omega)$.

7 An example where the Karhunen-Loève approach is inefficient: the ramp function

During one of the many wavelet conferences at Marseille I challenged the established Karhunen-Loève algorithm with a counter-example which was aimed to show that wavelets were superior. My example was the following process. We pick a point at random ω in $(0,1)$. Then the random ramp function we want to analyze is $f_\omega(t) = t$ if $0 \leq t < \omega$ while $f_\omega(t) = t - 1$ if $\omega \leq t < 1$. It is trivial to check that the corresponding Karhunen-Loève basis is the Fourier basis $\frac{\sqrt{2}}{2}\sin(\pi n t), 1 \leq n$. But each realization of this process is poorly approximated in this basis. The reason being that the Fourier coefficients c_n of $f_\omega(t)$ decay as $1/n$. If we want an L^2-norm error less than 10^{-3}, we need 10^6 terms in the expansion! In contrast when one is using a wavelet basis with two vanishing moments (for example the Alpert-Rokhlin wavelets), the error will be less than $2^{-N/2}$ with N terms only. The reason why these Alpert-Rokhlin wavelets are not selected as the Karhunen-Loève basis is due to the fact that the location of these few terms in the wavelet series expansion of $f_\omega(t)$ depends on ω. This amounts to saying that the wavelet coefficients need to be sorted in order to pick the N non trivial terms in the wavelet expansion. This finite set $F(\omega)$ is defined by the condition that the support of the wavelet $2^{j/2}\psi(2^j t - k)$ should contain ω.

Then David Donoho told me that I was cheating. I was challenging a linear approximation scheme with a non-linear one. In the Karhunen-Loève algorithm one is trying to find an optimal q-dimensional vector space V_q which provides a best average fit with the data set. In the second case one is using a q-dimensional manifold M_q defined as the collection of all linear combination of q wavelets whatever be their location and size. Then M_q is not a linear space. It is a curved manifold and this curvature is responsible for a much better approximation rate as q tends to infinity.

This ramp example can be made periodic and the corresponding process will then be stationary. It can also be generalized to two dimensions. However the approximation rates in the $2-D$ case are distinct from the $1-D$ example. If we want the L^2-norm of the error to be less than N^{-1}, the number of needed

wavelets is $O(N^2)$ while $O(N^4)$ terms in the Fourier expansion are necessary for achieving the same precision.

If an image is modeled by a random finite collection of smooth curves delimitating some objects, the preceding discussion shows why wavelets perform better than Fourier.

8 A second visit to $u + v$ image models

As it was already mentioned, a common feature in all $u + v$ models is that natural images $f(x)$ are decomposed into a sum $u(x) + v(x)$. The first component $u(x)$ is well structured and has a simple geometric description since it models the objects that are present in the image. The second component $v(x)$ both contains the textured parts and the noise.

These two components $u(x)$ and $v(x)$ cannot be viewed as orthonormal or independent and this decomposition is not unique. For example both $u(x)$ and $v(x)$ contain arbitrarily high frequencies and a simply minded filtering cannot separate $u(x)$ from $v(x)$. Such a filtering would certainly kill $v(x)$ but also erase the edges which are present in $u(x)$.

This $f = u+v$ modeling is reminiscent of approximation theory and more precisely it mimics the theory of interpolation. In this context, one wants to write a generic function $f(x)$ as a sum of a "good function" $u(x)$ which is more regular than $f(x)$ plus a "bad function" $v(x)$ which is small in some sense. An example is the celebrated Calderón-Zygmund decomposition of an L^1 function $f(x)$ into a sum of an L^2 function $u(x)$ plus an oscillating part $v(x)$ carried by a set with a small measure.

We now want to explain why in many $u+v$ image models, the $u(x)$ component is assumed to be a function with bounded variation. In the case of image processing, we want to detect objects delimitated by contours. Then these objects can be modeled by some planar domains $D_1, ..., D_n$ and the corresponding contours or edges will be modeled by their boundaries $\partial D_1, ..., \partial D_n$. In this model, the function $u(x)$ is assumed to be smooth inside $D_1, ..., D_n$ with jump discontinuities across the boundaries $\partial D_1, ..., \partial D_n$. However we do not want to break an image into too many pieces and the penalty for a domain decomposition of a given image will be the sum of the lengths of these edges $\partial D_1, ..., \partial D_n$. But this sum of lengths is approximatively one of the two terms which appear in the BV norm of $u(x)$. The BV norm of a

function $f(x)$ is defined as the total mass of the distributional gradient of $f(x)$ and we will return to this definition in the next section.

This discussion leads to a specific $u+v$ model where $u(x)$ will be assumed to belong to the space BV of functions with bounded variation and $v(x)$ will be measured by a simply minded energy criterion which says that $\|v\|_2$ is sufficiently small.

9 The space BV of functions with bounded variation in the plane

We first consider the general case where the dimension n is larger than 1. In the one-dimensional case, the space BV is trivial since it is isomorphic to the space of all bounded Radon measures on the line. Assuming $n \geq 2$, we say that a function $f(x)$ defined in R^n belongs to BV if (a) $f(x)$ vanishes at infinity in a weak sense and (b) the distributional gradient of $f(x)$ is a bounded Radon measure. The BV norm of f is denoted by $\|f\|_{BV}$ and defined as the total mass of the distributional gradient of $f(x)$. The condition at infinity reads: $f \star \varphi$ tends to 0 at infinity whenever φ is a function in the Schwartz class.

A second and equivalent definition reads the following:

Definition 1 *A function $f(x)$ belongs to $BV(R^n)$ if it vanishes at infinity in the weak sense and if there exists a constant C such that*

$$\int_{R^n} |f(x+y) - f(x)|dx \leq C|y| \tag{9.1}$$

for each $y \in R^n$.

From this second definition, it is immediately concluded that if a real valued fonction $f(x)$ belongs to BV, so do $f^+(x) = sup(f(x), 0)$ and $f^-(x) = sup(-f(x), 0)$. In other words, it is often sufficient to consider non-negative functions in BV.

From now on the discussion will be restricted to the two-dimensional case. Then the space $BV(R^2)$ is contained in $L^2(R^2)$. This fact and some improvements on this Sobolev embedding theorem depend on the coarea identity we need to state explicitely.

If E is a measurable set, then χ_E will denote the indicator function of E and we would like to compute the BV norm of this indicator function χ_E whenever it is finite. If E is a Jordan domain delimited by a rectifiable boundary Γ, then the BV norm of χ_E is the total length of Γ. However if E is an open set whose boundary is denoted by ∂E, the BV norm of χ_E is in general smaller than the 1-dimensional Hausdorff measure $H^1(\Gamma)$ of its boundary $\Gamma = \partial E$ since $\chi_E = \chi_F$ almost everywhere does not imply $H^1(\partial E) = H^1(\partial F)$. An obvious counter-example is $E = D$ where D is the unit disc and $F = D \setminus L$ where L is any radius of our disc. A more interesting counter-example is provided by a swiss-cheese open set Ω with the following properties: Ω is a countable union of open squares Q_j, its closure is the unit square C and its boundary $\partial E = C \setminus \Omega$ is a connected compact set K with a positive Lebesgue measure. However the BV norm of χ_Ω is just the sum of the legths of ∂Q_j. In order to explain these facts, De Giorgi defined the reduced boundary $\partial^* E$ of a measurable set E and proved that the BV norm of χ_E is the 1-dimensional Hausdorff measure of its reduced boundary.

With these new notations the coarea identity reads as follows.

Theorem 1 *Let $f(x)$ be a real valued measurable function defined on the plane and let t be a positive real number. Let us denote by Ω_t the measurable set in the plane defined by*

$$\Omega_t = \{x \mid f(x) > t\} \tag{9.2}$$

Let $\partial^*\Omega_t$ be the reduced boundary of Ω_t and $l(t)$ the 1-dimensional Hausdorff measure of $\partial^*\Omega_t$. Then one has

$$\|f\|_{BV} = \int_0^\infty l(t)dt \tag{9.3}$$

This identity needs to be completed with the following observation.

$$f(x) = lim_{m\to\infty}[\int_{-m}^{\infty} \chi_{\Omega_t}(x)dt - m] \tag{9.4}$$

which, together with the coarea identity shows that a BV function admits a remarkable atomic decomposition.Here χ_A denotes the indicator function of the set A.

Let us be more specific and assume for the sake of simplicity that $f(x)$ is a smooth positive function vanishing at infinity. The atoms $a_\Omega(x)$ which will

be used are indicator functions χ_Ω of bounded connected open sets Ω with rectifiable boundaries $\partial\Omega$ ($\chi_\Omega(x) = 1$ on $\Omega, \chi_\Omega(x) = 0$ elsewhere). We will further restrict the definition of an atom by assuming that Ω is connected and simply connected. We know that the BV norm of such an atom χ_Ω is the total length of the reduced boundary $\partial^*\Omega$. If $f(x)$ is a positive function in the Schwartz class, then the indicator function of $\Omega_t = \{x \mid f(x) > t\}$, $t > 0$, is not an atom in general but rather a series of atoms with disjoint supports. Finally (9.3) and (9.4) show that any function $f(x)$ in BV is a Bochner integral $\int_0^\infty a_t(x)dt$ of such atoms. More precisely (9.3) implies

$$\int_0^\infty \|a_t\|_{BV} dt = \|f\|_{BV} \tag{9.5}$$

In what follows, we will prove some embedding theorems for the Banach space BV. The proof is based on the atomic decomposition. Indeed we can limit the discussion to the weakly dense subclass of BV consisting of simple functions. The weak density of simple functions inside BV refers to the weak-star topology of BV. This topology makes sense since the Banach space BV is a dual space X^* of a functional Banach space X. Returning to simple functions, we have $f(x) = \sum_0^\infty c_j a_j(x)$ where the $a_j(x)$ are the above defined atoms, the sum is indeed finite and $\sum_0^\infty |c_j| \|a_j\|_{BV} = \|f\|_{BV}$.

For proving an embedding theorem of the type $BV \subset Z$ where Z is some Banach space, it suffices to check that there exists a constant C such that the Z-norm of any atom $a_\Omega(x)$ in BV does not exceed C times the length of $\partial^*\Omega$.

Using this scheme one easily shows that BV is contained in the Lorentz space $L^{2,1}(R^2)$ which is included in L^2. Embeddings of BV inside Besov spaces will be discussed in section 13.

10 A last visit to the $u + v$ models

We now return to the general $u + v$ model for image processing which was introduced in section 8. This model has several variants. Some of them are deterministic and some are stochastic. In the deterministic models, $v(x)$ is taking care of the textures which are present in the given image and in the stochastic models, $v(x)$ also contains an additive noise.

The first model is due to Osher and Rudin [23]. The corresponding mathematical problem reads the following: knowing that a given function $f(x)$ is

the sum $f(x) = u(x) + v(x)$ with explicit bounds on the BV norm of the unknown function $u(x)$ and on the L^2-norm of the unknown function $v(x)$, we want to recover these unknown functions $u(x)$ and $v(x)$. The explicit condition on u reads $\|u\|_{BV} \leq C$ and the one on $v(x)$ is $\|v\|_2 < \epsilon$. There is no uniqueness and some more conditions are needed to find u and v. For instance, an optimal decomposition can be defined as follows. We keep the constraint on the BV norm of u and try to minimize the L^2 norm of v. To prove existence and uniqueness for this optimal decomposition, it suffices to consider the closed subset K of $L^2(R^2)$ defined by $\|u\|_{BV} \leq C$ and to define the optimal u as the point in K which minimizes the L^2 distance to $f(x)$.

As it was already mentioned a second and related problem consists in finding a fast algorithm that would yield a sub-optimal decomposition $f(x) = g(x) + h(x)$ where the corresponding bounds for g and h might be enlarged by a fixed multiplicative amount. This problem is addressed in [5].

A third approach to the "good and bad function decomposition" is the following. Given a function $f(x)$ in $L^2(R^2)$, we want to solve the variational problem

$$\omega(\lambda) = \inf\{J(u) = \|u\|_{BV} + \lambda\|v\|_2; \quad f = u + v\} \tag{10.1}$$

The tuning given by the large factor $\lambda = \epsilon^{-1}$ implies that the L^2-norm of v should be of the order of magnitude of ϵ. It is clear that solving this variational problem yields a suboptimal decomposition of $f(x)$. The mathematician will be intested in relating the growth of $\omega(\lambda)$ as λ tends to infinity to some properties of the L^2 function $f(x)$. For instance the space of all functions $f(x)$ for which $\omega(\lambda) = 0(\lambda^\gamma)$ as λ tends to infinity will be characterized in section 13, theorem 8.

A fourth splitting algorithm was proposed by Mumford and Shah [21]. In this algorithm, the u component belongs to the subspace SBV of BV, which consists of functions in BV whose distributional gradient does not contain a singular diffuse measure. In other words, this distributional gradient $Gradu$ is the sum between an L^1 function and a measure carried by a one dimensional singular set K. Then the Mumford-Shah penalty on the $u(x)$ component is a sum between two terms. The first term is the one-dimensional Hausdorff measure of K. The second one is the square of the L^2 norm of the gradient of $u(x)$ calculated on the complement of this singular set K. The third term of the $J(u)$ functional is the square of the L^2 norm of $v(x)$ [19].

A fifth approach to the decomposition $u+v$ was proposed by DeVore and Lucier. They replaced the BV norm of u by a Besov norm [9] in $J(u)$ as defined by (10.1).

A sixth approach concerns de-noising. Here $u(x)$ is an unknown function in BV which satisfies $\|u\|_{BV} \leq C$ and we are given noisy data $f(x) = u(x) + v(x)$. The noise $v(x)$ is often assumed to be a gaussian white noise and everything is sampled on a fine grid. It means that the sampled noise is a sequence which is i.i.d. $N(0, \sigma)$. We plan to apply this model to situations in image processing for which a correct stochastic model for images has not yet been found. The noise is assumed to be explicitely known and all expectations are calculated with respect to the corresponding probability law. Then the problem consists in finding an estimator $\hat{u} = F(f)$ which minimizes $\sup\{E[\|\hat{u} - u\|_2^2]\}$ where the supremum is taken over the ball $\|u\|_{BV} \leq C$. This ball is modeling our knowledge about the signal $u(x)$. An estimator is here defined as a non linear mapping F from the functional space containing u to the one containing the estimator. The interested reader is referred to an outstanding paper by David Donoho and Iain Johnstone [12] where these matters are discussed.

We do not intend to solve these problems and rather refer the reader to the existing literature. Our goal is less ambitious and our modest task will consist in comparing a Fourier analysis of BV functions to a wavelet analysis in order to conclude that the latter is a much better one. For that purpose some basic facts on wavelets need to be reviewed. The reader who is familiar with wavelet analysis should skip the following section and jump to section 12.

11 Wavelet analysis vs. Fourier analysis: defining terms

The goal of this section is to remind the reader of some basic properties of wavelet expansions which will be needed in section 13.

It was clear from the very beginning of signal processing that plain Fourier analysis does not make any good sense for real life signals. Indeed performing a Fourier analysis means integrating a given function $f(t)$ against $\cos(\omega t + \phi)$ from minus infinity to plus infinity. In the case of speech signal processing,

this integration would imply that we have to wait until the end of a speech to begin a Fourier transformation. This is not compatible with real-time transmission. Returning to mathematics, one cannot perform a Fourier analysis on the domain of definition of a function. In several dimensions it often happens that a function $f(x)$ is only defined on an open domain Ω. One cannot perform a Fourier analysis of $f(x)$ on this domain Ω. If we extend the function $f(x)$ by 0 outside Ω, this would cause jump discontinuities across the boundary of Ω and a Fourier analysis of this artificial function might be more sensitive to these jump discontinuities than to the intrisic properties of $f(x)$ inside Ω.

Scientists with such a gigantic stature as John von Neumann, Dennis Gabor, Leon Brillouin, Eugene Wigner,...pioneered modern signal processing in the forties. They addressed this problem and advocated for a windowed Fourier analysis. A windowed Fourier analysis is using a sliding window which is denoted by $w(t)$. This window is compactly supported and smooth. This smoothness condition is crucial. Otherwise multiplying the signal by this window would seriously alter the high frequency information contained in the signal and instead of analyzing the signal we would analyze the window. This window $w(t)$ is now translated by τ and modulated by multiplications with $\exp(i\omega t)$. We then obtain the famous Gabor wavelets as $g_{\omega,\tau} = w(t-\tau)\exp(i\omega t)$. But Gabor, von Neumann and the other scientists in his group were not satisfied by this continuous Gabor wavelet analysis. Indeed it mapped a function $f(t)$ of a real variable t into a function $F(\omega,\tau)$ of two variables ω and τ. If a given signal is sampled over 10^5 points we are now dealing with 10^{10} points in the time-frequency plane. We do not want to waste our bit budget in such a foolish way. That is why the above mentioned pioneers wanted a discrete version of this Gabor wavelets analysis. Some heuristics about paving the time-frequency plane with Gibbs cells led to define the optimal sampling by $\tau = 2k\pi, \omega = j, \quad j,k \in Z$. But it was later proved by two other physicists (Francis Low and Roger Balian) that ther exist L^2 functions that cannot be decomposed into a convergent series of such Gabor wavelets. Finally everything was repared when Kenneth Wilson (Nobel prize winner in 1984) reshaped Gabor and produced orthonormal time-frequency atoms leading to fast algorithms for local Fourier analysis. The key idea of Wilson was to modify the modulation by $\exp(i\omega t)$ in a way that is reminiscent of the DCT algorithm. We will not yield further details and more details can be found in [15].

Time-scale algorithms and wavelet analysis can be defined as an alternative to the classical windowed Fourier analysis and to time-frequency analysis. In the latter case the goal is to measure the local frequency content of a signal while in the wavelet case one is comparing several magnifications of this signal, with distinct resolutions. These magnifications are often called "zoomings". The building blocks of a windowed Fourier analysis are sines and cosines (waves) multiplied by a sliding window. In a wavelet analysis the window is already oscillating and is called a mother wavelet. This mother wavelet $\psi(t)$ has a compact support (or a rapid decay at infinity), is smooth and satisfies the fundamental condition $\int_{-\infty}^{+\infty} \psi(t)\, dt = 0$, which means that in some weak sense $\psi(t)$ is oscillating. Moreover the mother wavelet should also satisfy a technical but crucial condition which will be revealed soon. The mother wavelet is no longer multiplied by sines or cosines. Instead it is translated and dilated by arbitrary translations and dilations. That is the way the mother wavelet $\psi(t)$ generates the other wavelets $\psi_{a,b}(t) = a^{-1/2}\psi((t-b)/a)$ (where $a > 0$, $-\infty < b < +\infty$) which are the building blocks of a wavelet analysis. The parameter a measures the average width of the wavelet $\psi_{a,b}(t)$ while the parameter b gives the position. These dilations (by $1/a$) are precisely the magnifications we alluded to. The wavelet coefficients of a function $f(t)$ of the real variable t are the scalar products $W(a,b) =< f, \psi_{a,b} >$ (where $a > 0$ and $-\infty < b < +\infty$). Here and in what follows $< u, v >= \int_{-\infty}^{+\infty} u\overline{v}\, dt$. In other words one is computing the correlations between the function to be analyzed and translated and/or dilated versions of the analyzing wavelet ψ. The original function $f(t)$ can always be recovered as a linear combination of these wavelets $\psi_{a,b}(t)$ and, up to a normalization which will be specified, the coefficients of this combination are precisely the wavelet coefficients $W(a,b)$.

A wavelet analysis is either continuous, semi-discrete, orthonormal or biorthogonal. In the first case one is using Calderón's reproducing identity. For the reader's convenience, let us stress the relationship between this identity and a wavelet analysis and take this opportunity for giving the precise definition of a wavelet. We now consider the n-dimensional case. An analyzing wavelet $\psi(x)$ is a function which satisfies the condition

$$\int_0^\infty |\hat{\psi}(t\xi)|^2\, dt/t = 1 \tag{11.1}$$

for almost every ξ in R^n. This condition is named "admissibilty". Then the continuous wavelet coefficients of a function $f(x)$ in $L^2(R^n)$ are defined as

$F(x,t) =< f, \psi_{x,t} >$ where $\psi_{x,t}(y) = t^{-n/2}\psi((y-x)/t)$. For recovering $f(x)$ it suffices to c ombine all these wavelets $\psi_{x,t}$ with precisely these coefficients. In other terms, we obtain

$$f(x) = \int_0^\infty \int_{R^n} F(y,t)\psi_{y,t}(x)dy\,dt/t^{n+1} \tag{11.2}$$

which is exactly Calderón's reproducing identity. The relevance of a continuous wavelet analysis will heavily depend on the properties of the analyzing wavelet $\psi(x)$. Two usual choices are the Morlet wavelet (a modulated gaussian which does not exactly satisfy the admissibility requirement) or the mexican hat (the second derivative of a gaussian).

What Calderón's reproducing identity tells us is the following: a wavelet analysis gives a recipe for (a) measuring the local fluctuation coefficients of a given function f, around any point x, at any scale t and for (b) reconstructing f with all these fluctuation coefficients. In other terms at any given scale $a > 0$, f is decomposed into the sum of a trend at the scale a and of a fluctuation around this trend. The trend is given by the contribution of scales $t > a$ in Calderón's reproducing identity and the fluctuation is given by the scales $t < a$.

Let us now return to the one-dimensional case and study orthonormal wavelet bases. A "mother wavelet" will now be defined as a function $\psi(t)$ enjoying the three following properties:

(11.3) $\psi(t)$ is a smooth function (with $r-1$ continuous derivatives and a bounded derivative of order r)

(11.4) $\psi(t)$ together with its derivatives of order less than r has a rapid decay at infinity

(11.5) the collection $\psi_{j,k}(t)$ defined by $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$, $j,k \in Z$, is an orthonormal basis for $L^2(R)$.

The first problem in the theory is to construct such functions $\psi(t)$ and the second one is to show that the wavelet coefficients yield a relevant information. The very first example of a mother wavelet was given by A.Haar in 1909. The Haar wavelet $h(t)$ is defined by $h(t) = 1$ on $[0, 1/2)$, $h(t) = -1$ on $[1/2, 1)$ and $h(x) = 0$ elsewhere. In that case $r = 0$.

But about eighty years were needed until Ingrid Daubechies proved that for each $r \geq 1$, one can construct a function $\psi(t)$ of class C^r with compact

support and satisfying the above conditions (11.3) and (11.5), the second condition being obvious [6]. A detour with a visit to the signal processing community and a reshaping of the subband coding algorithms were needed to build these Daubechies wavelets. Today we know that this detour is absolutely necessary. Let us provide the reader with a leisurely description of this detour. It begins with the definition of a multiresolution analysis.

A multiresolution analysis of $L^2(R)$ is a ladder V_j, $j \in Z$, of closed subspaces of $L^2(R)$ enjoying the following four properties:

(11.6) the intersection $\cap V_j$, $j \in Z$, is reduced to $\{0\}$,

(11.7) the union $\cup V_j, j \in Z$, is dense in $L^2(R)$,

(11.8) $f(t)$ belongs to V_j if and only if $f(2t)$ belongs to V_{j+1} and finally

(11.9) there exists a smooth and localized function $f(t)$ such that the collection $\varphi(t-k), k \in Z$, be an orthonormal basis for V_0. This function $\varphi(t)$ is named the "scaling function".

Multiresolution analysis is a natural concept for people working on splines since refinements of meshes provide trivial examples. The relation between our wavelet basis and a multiresolution analysis is given by the condition that $\psi(t-k), k \in Z$, is an orthonormal basis of the orthogonal complement W_0 of V_0 in V_1. By an obvious rescaling one obtains the fact that $2^{j/2}\psi(2^j t - k), k \in Z$, is an orthonormal basis for the orthonormal complement W_j of V_j into V_{j+1}. It is then clear that the full collection $\psi_{j,k}$ is an orthonormal basis for $L^2(R)$. In this construction the "mother wavelet" is built from the scaling function $\varphi(t)$. The converse problem consists in asking whether any orthonormal wavelet basis $2^{j/2}\psi(2^j t - k)$, $j \in Z, k \in Z$, can always be constructed by this procedure. This cannot be true in general, as a counterexample due to J.L.Journé shows. However if $\psi(t)$ satisfies some reasonable smoothness and localization properties, P.G.Lemarié-Rieusset proved that an orthormal wavelet basis is always coming from a multiresolution analysis [13].

Multiresolution analysis is a mathematical concept which highlighted pyramidal algorithms and subband coding. This dictionary between mathematical analysis and signal processing was elaborated by Stéphane Mallat and the author [14], [16]. Let us provide the reader with a few entries from this dictionary.

First the quadrature mirror filters used in subband coding give the matrix representation of this orthonormal decomposition $V_{j+1} = V_j \oplus W_j$.

Similarly the pyramidal algorithms which were used in image processing and multiresolution analysis of $L^2(R^2)$ are closely related concepts. If $\Gamma_j, j \in Z$, are the sampling grids, then the sampling operator $P_j : L^2(R^2) \rightarrow l^2(\Gamma_j)$ is nothing else but the orthogonal projection from $L^2(R^2)$ onto V_j. To clarify this correspondence, it suffices to associate each vector of the $2-D$ orthonormal basis $2^j\varphi(2^j x - k), k \in Z^2$ of V_j to the corresponding point $k2^{-j}$ of the grid Γ_j. This is a natural choice since this scaling function is centered around this point. It will provide an isometrical isomorphism between V_j and $l^2(\Gamma_j)$. The "coarse to fine" algorithm in the pyramidal algorithm reflects the canonical embedding of V_j inside V_{j+1} while the "fine to coarse" algorithm corresponds to the orthogonal projection from V_{j+1} onto V_j.

The multiresolution analysis used by Ingrid Daubechies to construct her wavelets is highly non standard and was never considered by "spline people" who were unaware of the work achieved by signal processing researchers on quadrature mirror filters. Moreover spline specialists were not interested in the spaces W_j giving the missing details needed for a "coarse to fine algorithm".

As it was already mentioned, this connection between wavelet analysis, multiresolution analysis, pyramidal algorithms and subband coding was first stressed by S.Mallat [10] and the filters which produce the Daubechies wavelet $\psi(x)$ by a cascade algorithm were not unfamiliar to the signal processing community. The reason why they did not discover these Daubechies wavelets seems due to the fact that they did not know that any signal could be decomposed into a sum of wiggling waveforms which are obtained by dilations and translations from a mother wavelet $\psi(x)$. This idea however was quite familiar to mathematicians around Guido Weiss who created the so-called atomic decompositions of the Hardy space $H^1(R^n)$. However most of the atoms used by Guido Weiss do not satisfy the smoothness conditions we imposed on wavelets [16]. Strikingly all the ingredients needed to build Daubechies orthonormal wavelet bases were available in separate places of science or technology. To my opinion the "wavelet wisdom" developed by A.Grossmann and J.Morlet was the essential unifying concept that helped the construction of these remarkable orthonormal bases.

One cannot expect any serious understanding of what wavelet analysis means without a deep knowledge of the corresponding operator theory.

Indeed there are several interesting choices of orthonormal wavelet bases and one needs to know if some results obtained by using one specific basis would still be true with an other one. For answering this problem one needs a dictionary between all those bases. This dictionary is provided by the Calderón-Zygmund theory. Indeed unitary operators that map one orthonormal wavelet basis into an other one are Calderón-Zygmund operators. One key ingredient in this operator theory is the ability of rescaling global L^2-estimates for obtaining pointwise information. Therefore, as stressed by E.Stein, the group of dilations plays a crucial role in the Calderón-Zygmund theory. The Hilbert transform H, defined by

$$H(f)(x) = p.v. \int_{-\infty}^{+\infty} f(x-y)\, dy/y \tag{11.10}$$

is the prototype of a Calderón-Zygmund operator. It is the only non-trivial operator which is translation and dilation invariant (only positive dilations are considered). If $\psi(x)$ is a mother wavelet generating an orthonormal wavelet basis, its Hilbert transform $H(\psi) = w$ has the same property and H maps the orthonormal basis $\psi_{j,k}$ onto the orthonormal basis $w_{j,k}$. Therefore any information obtained by inspecting coefficients in some orthonormal wavelet expansion is necessarily invariant under the action of the Hilbert transform. For example, one cannot decide if a function is continuous by inspecting its wavelet coefficients. Indeed continuity is not preserved by the Hilbert transformation. Similarly it is impossible to obtain a wavelet coefficients based criterion for deciding if a given distribution is a measure. Finally similar remarks apply to the $2-D$ case and the space BV cannot be characterized by size estimates on the wavelet coefficients since BV is not preserved by Calderón-Zygmund singular integral operators. The Beurling transformation B is defined by $B(\frac{\partial f}{\partial z}) = \frac{\partial f}{\partial \bar{z}}$, $z = x+iy$. Then B is a Calderón-Zygmund singular integral operator which is unitary on L^2. Aline Bonami and S Poornima proved the following theorem [1]:

Theorem 2 . *The Beurling transformation B does not map BV into itself.*

Theorem 2 is given an interesting perspective if one returns to the definition of the Hardy space $H^1(R)$ which is the collection of all functions f in L^1 whose Hilbert transform $H(f)$ also belongs to L^1. Similarly, as it will be later explained, the space BV, up to a trivial isomorphism, is the space of

all real valuedfunctions g such that both g and $B(g)$ are bounded Radon measures. The Hilbert transformation H acts continuously on H^1 and it was not unreasonable to conjecture that B mapped BV into itself. For a better understanding of the relation between BV and the Beurling transformation, one starts with a function f in BV and writes $u = \partial f/\partial x, v = \partial f/\partial y$. Then $g = u - iv = B(h)$ where $h = u + iv$. Therefore g and h are bounded Radon measures and the converse statement is just as easy: if both $\partial f/\partial z$ and $\partial f/\partial \overline{z}$ are bounded Radon measures, then f belongs to BV.

Corollary. *The space BV cannot be characterized by size properties on wavelet coefficients.*

Indeed the Beurling transformation is a unitary Calderón-Zygmund operator acting on L^2. Therefore B maps any orthonormal wavelet basis $\psi_{j,k}$ onto an other one $w_{j,k}$. Moreover if ψ belongs to the Schartz class, then all its moments vanish and w also belongs to the Schwartz class. If the following statement was true:

Statement. *Whenever $\psi_{j,k}$ is an orthonormal wavelet basis where ψ belongs to the Schwartz class, then BV is characterized by size conditions on the corresponding wavelet coefficients $< f, \psi_{j,k} >$.*

then we would deduce that BV is invariant under the action of the Beurling transformation. This is not the case and the above statement was wrong.

However it will be later shown (theorem 7) that BV is almost characterized by such size estimates. The Banach space of these wavelet coefficients is sitting somewhere between l^1 and weak-l^1.

On the opposite, Hölder classes as well as the two-microlocal spaces are preserved by the Hilbert transformation which means that Hölder exponents can be computed through inspecting a wavelet expansion. This explains why wavelets are playing a key role in the so-called multifractal signal processing which relies on the computation of local Hölder exponents [17].

This section will end with describing the two-dimensional wavelets which will be used next. For constructing these $2-D$ wavelets, we both need the one-dimensional wavelet ψ and the corresponding scaling function φ. Then the three $2-D$ mother wavelets are

(a) $\psi_1(x_1, x_2) = \varphi(x_1)\psi(x_2)$,

(b) $\psi_2(x_1, x_2) = \psi(x_1)\varphi(x_2)$ and finally

(c) $\psi_3(x_1, x_2) = \psi(x_1)\psi(x_2)$.

The collection of these three $2-D$ wavelets will be denoted by F and the $2-D$ wavelet analysis is described by the following theorem:

Theorem 3 *For each positive exponent r, there exist three functions $\psi_m, m = 1,2,3$, with the following properties: each $\psi_m(x_1, x_2)$ is compactly supported and belongs to the Hölder space C^r* (11.11)
$2^j\psi_m(2^j x_1 - k_1, 2^j x_2 - k_2), j \in Z, k = (k_1, k_2) \in Z^2, m = 1,2,3$, is an orthonormal basis for $L^2(R^2)$ (11.12).

This theorem will be used in the next section and for simplifying the notation, we will write $\lambda = (j,k,m)$ and denote by $\psi_\lambda(x)$ the corresponding wavelet. Then λ belongs to $\Lambda = Z \times Z^2 \times \{1,2,3\}$.

12 Quantization issues: Fourier series vs. wavelet series

Whenever a computer is used, the true coefficients arising in some expansion will be replaced by approximations to a given precision. What happens to the expansion after a quantization is performed is a problem to be addressed. Some representations are more sensitive than others to quantization. Wavelet expansions have the advantage that the effect of small changes over the coefficients will only have a local influence. This fact is related to the following observation. If $\psi_\lambda, \lambda \in \Lambda$, is an orthonormal wavelet basis and if $m_\lambda, \lambda \in \Lambda$, is a multiplier sequence which is used to shrink the wavelet coefficients in this basis, then the corresponding multiplier operator M defined by $M(\psi_\lambda) = m_\lambda\psi_\lambda$, is a Calderón-Zygmund operator.

When the trigonometric system is used, any change on any coefficient will affect the resulting function globally.

More precisely let us study the non-linear mapping $Q_\epsilon(f)$ which is defined on 2π periodic functions by the following algorithm. We start with the Fourier series expansion of a 2π-periodic function $f(x)$ and replace by 0 all the coefficients whose absolute value is less than ϵ. We then obtain f_ϵ and write $f_\epsilon = Q_\epsilon(f)$. We would like to understand the behavior of this operator as ϵ tends to 0. The following theorem easily follows from the construction of Rudin-Shapiro polynomials.

Theorem 4 *For each exponent α less than $1/2$ there exists a 2π-periodic function $f(x) = f^{\alpha}(x)$ belonging to the Hölder space C^{α} such that the L^{∞} norm of $Q_{\epsilon}(f) = f_{\epsilon}$ tends to infinity as ϵ tends to 0. More precisely*

$$\|f_{\epsilon}\|_{\infty} > C\epsilon^{-\beta} \tag{12.1}$$

where $C = C(f)$ is a positive constant and β which is defined by

$$\beta = \frac{1-2\alpha}{1+2\alpha}$$

is also positive.

When α is larger than $1/2$ the Fourier coefficients of our function belong to l^1 and $\|f_{\epsilon}\|_{\infty} < C$ for some constant C.

The blow-up described by theorem 4 cannot occur with wavelet expansions. Indeed the Hölder space C^{α} is characterized by size conditions on the wavelet coefficients.

13 Fourier series vs. wavelet series: expansions of BV functions

For the sake of simplicity, let us first study periodic functions in BV. Let $f(x_1, x_2)$ be a function of two real variables which is 2π-periodic in each variable. We then abbreviate in saying that $f(x)$ is 2π-periodic. Let us write the Fourier series of $f(x)$ as $f(x) = \sum_{k \in Z^2} c(k_1, k_2)\exp(ik.x)$ with $k = (k_1, k_2)$. Let us assume that $f(x)$ belongs to BV on $[0, 2\pi]^2$. Then we already know that $c(k)$ belongs to l^2. For such functions, Jean Bourgain proved the following

Theorem 5 *There exists a constant C such that for any 2π-periodic function $f(x)$ in $BV(R^2)$, we have*

$$\sum |c(k)|(|k|+1)^{-1} \leq C\|f\|_{BV} \tag{13.1}$$

This estimate complements $\sum |c(k)|^2 < \infty$ and these two results obviously follow from a sharper estimate given by

$$\sum_{j=0}^{\infty} s_j \leq C\|f\|_{BV} \tag{13.2}$$

where $s_j = (\sum_{2^j \leq |k| < 2^{j+1}} |c(k)|^2)^{1/2}$.

This is a mixed $l^1(l^2)$ estimate on Fourier coefficients of a BV function. It is optimal in the sense that there exists a function in BV for which $\sum |c(k)|^p = \infty$ for any $p < 2$. An example is given by $f(x) = |x|^{-1}(log|x|)^{-2}\varphi(x)$ where $\varphi(x)$ is any smooth function which vanishes when $|x| > 1/2$ and is identically 1 around the origin. Then the Fourier coefficients $c(k)$ of $f(x)$ can be estimated by $|c(k)| \simeq |k|^{-1}(log|k|)^{-2}$ which obviously implies $\sum |c(k)|^p = \infty$ as announced. The sorted Fourier coefficients of this function behave as $n^{-1/2}(logn)^{-2}$. This counter-example shows that nothing better than l^2 can be expected inside the dyadic blocks of the Fourier series expansion of a funcion $f(x)$ in BV.

Now this estimate (13.2) can be rewritten as a Besov norm estimate. Indeed let $\Delta_j(f)$ denote the dyadic blocks of the Fourier series expansion of $f(x)$. For defining $\Delta_j(f)$ we only retain the frequencies $k \in \Gamma_j$ in the Fourier expansion of f where Γ_j is the dyadic annulus defined as $\{k\,|2^j < |k| \leq 2^{j+1}\}$. We then obviously have $f(x) = c_0 + \sum_0^\infty \Delta_j(f)$ and our next theorem reads:

$$\sum_{0}^{\infty} \|\Delta_j(f)\|_2 \leq C\|f\|_{BV} \tag{13.3}$$

This theorem will be further improved. This improved version is not using a Fourier series expansion any more and we can therefore give up the periodic setting and switch to the space $BV(R^2)$ and to a Littlewood-Paley analysis.

There are two approaches to Littlewood-Paley expansions, as it was the case for a wavelet analysis. We may start with a compactly supported smooth function ψ with enough vanishing moments such that the Fourier transform Ψ of ψ satisfies $\sum_0^\infty \Psi(2^{-j}\xi) = 1$ whenever $|\xi| > 1$. Next we write $\psi_j = 2^{2j}\psi(2^j x)$. Finally $\Delta_j(f)$ is the convolution product $f * \psi_j$. Then there exists a smooth and compactly supported function φ whose Fourier transform Φ satisfies $\Phi(\xi) + \sum_0^\infty \Psi(2^{-j}\xi) = 1$ identically. We denote by $S_0(f)$ the convolution product between f and φ and we have $S(f) + \sum_0^\infty \Delta_j(f) = f$. But we may directly start with a sufficiently large integer N and a smooth compactly supported function $\varphi(x)$ such that $\int_{R^2} \varphi(x)\,dx = 1$ and $\int_{R^2} x^\alpha \varphi(x)\,dx = 0$ for

$1 \leq |\alpha| \leq N$. Next we define $\psi(x) = 4\varphi(2x) - \varphi(x)$ and the dyadic blocks will be the convolution products $\Delta_j(f) = \psi_j * f$ where $\psi_j(x) = 4^j\psi(2^j x)$. In these two approaches ψ is playing the role of a wavelet and φ of the scaling function. The value of the integer N is related to the amount of smoothness one wants to analyze.

With these notations (13.3) can be generalized to all exponents p in $(1, 2]$. Indeed one has (Y.M.)

Theorem 6 *There exists a constant C such that for every function f in $BV(R^2)$, and for every exponent p with $1 < p \leq 2$, we have*

$$\sum_{-\infty}^{+\infty} 2^{js}\|\Delta_j(f)\|_p \leq C_p\|f\|_{BV} \tag{13.4}$$

with $s = -1 + (2/p)$ and $C_p \leq C/(p-1)$.

The proof of this theorem is quite simple. Let us first observe that, up to a multiplicative constant, the left hand side of (13.4) does not depend on the definition of the Littlewood-Paley expansions into dyadic blocks. This remark will permit to use a compactly supported ψ in the definition of the Littlewood-Paley analysis which is used. Using the atomic decomposition of BV, it suffices to prove the theorem for an individual atom. If L is the length of $\partial\Omega$, we are led to distinguish between $j \leq q$ and $j > q$ where 2^{-q} is of the order of magnitude of L. When $j \leq q$, we use the support compact of ψ and can uniformy bound $|\Delta_j(f)|$ by $4^j\text{vol}(\Omega)$ if the distance from x to Ω does not exceed $C2^{-j}$ and by 0 if this distance exceeds $C2^{-j}$. For each such exponent j, the L^p-norm we need to estimate is easily bounded by $C2^{2j(1-1/p)}\text{vol}(\Omega)$ and the sum over j does not exceed $C_pL^{-1}\text{vol}(\Omega)$ which is less than C'_pL (by the isoperimetric inequality). When j is larger than q, we take in account the cancellation of the analyzing wavelet ψ and observe that $\Delta_j(f)$ vanishes unless the distance from x to $\partial\Omega$ is less than 2^{-j}. In the latter case $|\Delta_j(f)(x)| \leq C$. Since $\partial\Omega$ is rectifiable, the area of this set of point is less than $C2^{-j}L$. Finally this yields an L^p norm $\|\Delta_j(f)\|_p$ not exceeding $C(2^{-j}L)^{1/p}$. Summing over $j < q$ again yields $C'L$.

Corollary 1 *If $f(x)$ belongs to $BV(R^2)$, and if $\psi_{j,k}(x) = 2^j\psi(2^jx - k)$, $j \in Z, k \in Z^2$, is an orthonormal wavelet basis of $L^2(R^2)$ where ψ is smooth*

and localized, then the corresponding wavelet coefficients $c(j,k) =< f, \psi_{j,k} >$ *satisfy*

$$\sum_j \{\sum_k |c(j,k)|^p\}^{1/p}\} \leq C/(p-1)\|f\|_{BV}, \quad 1 < p < 2 \tag{13.5}$$

Corollary 2 *With the same notations as above, we have*

$$\{\sum_j \sum_k |c(j,k)|^p\}^{1/p} \leq C/(p-1)\|f\|_{BV} \tag{13.6}$$

Corollary 3 *With the same notations, let us assume* $\|f\|_{BV} \leq 1$. *For each integer* m, *let* N_m *be the cardinality of the set on indices* (j,k) *such that* $|c(j,k)| > 2^{-m}$. *Then*

$$N_0 + ... + 2^{-m} N_m \leq C(m+1) \tag{13.7}$$

It means that for most m's we have $N_m \leq C2^m$ since the average of $2^{-m}N_m$ is 0(1).

Indeed one has $N_m \leq C2^m$ for all m. Keeping the notation of theorem 3, the sharp estimate $N_m \leq C2^m$ will be rephrased in the following theorem (A.Cohen, Y.M. and F.Oru):

Theorem 7 *Let* $\psi_\lambda, \lambda \in \Lambda$, *be a two-dimensional orthonormal wavelet basis as described in Theorem 3. Then for every* f *in* $BV(R^2)$, *the wavelet coefficients* $c_\lambda =< f, \psi_\lambda >, \lambda \in \Lambda$ *belong to weak* $l^1(\Lambda)$.

This theorem was proved by A.Cohen et al. [5] in the Haar system case. The general case was obtained by the author and the best reference is [22].

In other words, if $c_\lambda =< f, \psi_\lambda >$ and if the $|c_\lambda|, \lambda \in \Lambda$, are sorted out by decreasing size, we obtain a non-increasing sequence c_n^* which satifies $c_n^* \leq C/n$ for $1 \leq n$.

One cannot replace the vector space weak-$l^1(\Lambda)$ by $l^1(\Lambda)$ in theorem 7. Indeed let $f(x)$ be the indicator function of any smooth domain Ω and let L be the length of the boundary of Ω. Then when $2^j L > 1$, the cardinality of the set of λ such that $2^{-j} < |c_\lambda| \leq 2^{-j+1}$ is precisely $2^j L$. This does not mean that theorem 7 is optimal. Indeed BV is a Banach space which is

the dual X^* of a separable Banach space X while weak l^1 does not have this property. Therefore BV and weak l^1 cannot be isomorphic. A second remark concerns (13.5). This statement is not implied by the weak l^1 property. A last observation is the obvious remark that functions in BV cannot be characterized by size estimates on wavelet coefficients [1]. But theorem 7 and (13.5) show that BV is almost characterized by such estimates. The vector space Y of wavelet coefficients of BV functions is sitting somewhere between l^1 and weak l^1. We can now return to the problem raised in section 10. Let γ be an exponent in $(0,1)$. We want to characterize the space of all functions $f(x)$ in $L^2(R^2)$ such that

$$\omega(\lambda) = inf\{J(u) = \|u\|_{BV} + \lambda\|v\|_2; f = u + v\} = 0(\lambda^\gamma), \lambda \to \infty \quad (13.8).$$

We then obtain

Theorem 8 *We have $\omega(\lambda) = 0(\lambda^\gamma)$ when λ tends to infinity if and only if the sorted wavelet coefficients of $f(x)$ satisfy $c_n^* = 0(n^{-\alpha})$ where $\alpha = 1 - \gamma/2$.*

14 Improved Gagliardo-Nirenberg inequalities

We now want to relate theorem 7 with some improved Gagliardo-Nirenberg inequalities. Let us start with the Sobolev embedding of BV into $L^2(R^2)$.

The estimate

$$\|f\|_2 \leq C\|f\|_{BV} \quad (14.1).$$

is obviously consistent with translations and dilations. Indeed, for any positive a and $f_a(x) = af(ax)$, we will have $\|f_a\|_2 = \|f\|_2$ and similarly $\|f_a\|_{BV} = \|f\|_{BV}$. But (14.1) is not consistent with modulations: if M_ω denotes the pointwise multiplication operator with $\exp(i\omega x)$, then M_ω acts isometrically on L^2 while $\|M_\omega f\|_{BV}$ blows up as $|\omega|$ when $|\omega|$ tends to infinity.

For addressing this invariance through modulations, let us introduce an adapted Besov norm.

Definition 2 *Let B be the Banach space of all tempered distributions $f(x)$ for which there exists a constant C such that for $g(x) = exp(-|x|^2)$ and $g_{a,b} = ag(a(x-b))$, the following condition is satisfied:*
there exists a constant C such that for any $a > 0, b \in R^2$, we have

$$| < f, g_{a,b} > | < C \tag{14.2}$$

The infimum of these constants C is the norm of f in B and is denoted by $\|f\|_\epsilon$.

It is easily proved that this Banach space coincides with the space of second derivatives of functions in the Zygmund class. Therefore B is the homogeneous Besov space $B_\infty^{-1,\infty}$ of regularity index -1.

We then have

Theorem 9 *There exists a constant C such that for any f in $BV(R^2)$ we have*

$$\|f\|_2 \leq C[\|f\|_{BV}\|f\|_\epsilon]^{1/2} \tag{14.3}$$

and $\|f\|_\epsilon$ is the weakest norm obeying the same scaling laws as the L^2 or BV norm for which (14.3) is valid.

To better understand this theorem, let us stress that we always have $\|f\|_\epsilon \leq \|f\|_{BV}$ and the ratio $\|f\|_\epsilon/\|f\|_{BV}$ between these norms is denoted by β and is expected to be small in general. Then (14.3) reads

$$\|f\|_2 \leq C\beta^{1/2}\|f\|_{BV} \tag{14.4}$$

which yields a sharp estimate of the ratio between the L^2 norm and the BV norm of f. Moreover $\beta^{1/2}$ in (14.4) is sharp as the example of $f(x)$=exp$(i\omega x)w(x)$ shows. Indeed if $|\omega|$ tends to infinity and $w(x)$ belongs to the Schwartz class, then $\|f\|_2$ is constant, $\|f\|_\epsilon \simeq |\omega|^{-1}\|f\|_\infty$, and finally $\|f\|_{BV} \simeq |\omega|\|w\|_1$. In this example β is of the order of magnotude of $|\omega|^{-2}$ which corresponds to $\beta^{1/2} \simeq |\omega|^{-1}$.

The proof of (14.3) is straightforward. One uses the following trivial estimate on sequences

$$\sum_{n=1}^{\infty} |c_n|^2 \leq 2\|c_n\|_\infty \|nc_n\|_\infty \tag{14.5}.$$

Then one applies theorem 7 to an orthonormal wavelet basis of class C^2. If c_n^* denotes the non increasing rearrangement of the wavelet coefficients $|c(\lambda)|, \lambda \in \Lambda$, then $\|nc_n^*\|_\infty$ is precisely the norm of $c(\lambda)$ in the space weak l^1.

Let us observe that (14.3) is an interesting improvement on the celebrated Gagliardo-Nirenberg estimates. These estimates read in the two-dimensional case

$$\|D^j f\|_p \leq C(\|D^m f\|_r)^\sigma (\|f\|_q)^{1-\sigma} \tag{14.6}$$

where $1 < p, q, r < \infty, j/m < \sigma < 1$ and $1/p - j/2 = \sigma(1/r - m/2) + (1-\sigma)/q$.

The notation $\|D^j f\|_p$ means $sup\{\|\partial^\alpha f\|_p; |\alpha| = j\}$. For comparing our new estimate to the Gagliardo-Nirenberg estimate (14.6), we will assume $m = 2, j = 1, p = 2$ and $r = 1$. This either implies $s = 1$ or $q = \infty$. In the first case, (14.6) easily follows from the embedding of BV into L^2 while in the second one (14.3) is an improvement on the Gagliardo-Nirenberg estimate since the L^∞ norm is replaced by a weaker one.

Theorem 9 generalizes to any dimension $n > 2$. It then reads

$$\|f\|_{n/n-1} \leq C(\|f\|_{BV})^{(1-1/n)} (\|f\|_\alpha)^{1/n} \tag{14.7}$$

where $\|f\|_\alpha$ is now defined as the optimal constant C for which one has $| < f, g_{a,b} > | \leq C$ with $g_{a,b} = ag(a(x-b)), a > 0, b \in R^n$, and $g(x) = exp(-|x|^2)$. In other words $\|f\|_\alpha$ is the norm of f in the homogeneous Besov space $B_\infty^{-(n-1),\infty}$.

Returning to L^2 norms, Albert Cohen, Wolfgang Dahmen, Ingrid Daubechies and Ron DeVore proved the following theorem (oral communication, still unpublished):

Theorem 10 *In any dimension $n \geq 1$, let us assume that a function f both belongs to BV and to the homogeneous space $B_\infty^{-1,\infty}$. Then we have*

$$\|f\|_2 \leq C(\|f\|_{BV} \|f\|_\epsilon)^{1/2} \tag{14.8}$$

where $\|f\|_\epsilon$ is norm of f in the Besov space $B_\infty^{-1,\infty}$.

The Besov norm of f can be defined as the optimal constant C for which one has $| < f, g_{a,b} > | \leq Ca^{2-n}, a > 0, b \in R^n$. Let us observe that BV is contained in L^2 if and only if $n = 2$. In other words when $n = 1$ or $n > 2$, the assumption $f \in B_\infty^{-1,\infty}$ complements $f \in BV$ and both are needed to get an L^2 estimate.

The proof of this theorem requires new estimates on wavelet coefficients of BV functions which are sharpening theorem 7. Indeed the above mentioned authors proved the following:

Theorem 11 *In any dimension $n \geq 1$, let us assume $\gamma < n-1$ where γ is a real exponent. Then for $f \in BV(R^n)$ and $\lambda > 0$, one has*

$$\sum_{\{|c(j,k)|>\lambda 2^{-j\gamma}\}} 2^{-j\gamma} \leq C\frac{\|f\|_{BV}}{\lambda} \tag{14.9}$$

where $c(j,k) = \int_{R^n} f(x)2^j\psi(2^jx-k)\,dx$

It is easily seen that this estimate is false when $\gamma = n-1$ and it is not difficult to construct sequences $c(j,k)$ belonging to weak-l^1 for which (14.9) is not fulfilled.

15 Improved Poincaré estimates

Here is a remarkable improvements on the standard Poincaré inequality. The standard Poincaré inequality reads as follows: If Ω is a connected bounded open set in the plane with a Lipschitz boundary $\partial\Omega$, then there exists a constant $C = C_\Omega$ such that for every f in $BV(\Omega)$ we have

$$\int_\Omega |f(x) - m_\Omega(f)|^2\,dx \leq C\|f\|^2_{BV} \tag{15.1}.$$

Here $m_\Omega(f)$ denotes the mean value of the function f over Ω. Such an estimate cannot be true in R^n for $n > 2$ since BV is not locally embedded in L^2 if $n \neq 2$. However there exists a sharpening of the Poincaré inequality which is valid in any dimension. Let $C^{-1}(\Omega)$ denote the Banach space of all distributions f on Ω which can be written as $f = \Delta F$ where F is the restriction to Ω of a function G belonging to the Zygmund class on R^n. The Zygmund class is defined by the classical condition that a constant C should exist such that

$$|G(x+y) + G(x-y) - 2G(x)| \leq C|y| \quad x, y \in R^n \tag{15.2}.$$

The norm of f in $C^{-1}(\Omega)$ is denoted by $\|f\|_\epsilon$ and is defined as the infimum of these constants C. This infimum is computed over all extensions G of F such that $f = \Delta F$ on Ω.

From now on Ω is assumed to have a smooth boundary $\partial\Omega$.

The improvement we have in mind is valid in any dimension and reads as follows:

Theorem 12 *With the preceding notations there exists a constant $C = C_\Omega$ such that*

$$\int_\Omega |f(x) - m_\Omega(f)|^2 \, dx \leq C\|f\|_{BV}\|f\|_{\cdot} \qquad (15.3).$$

This can be deduced from Theorem 10. Indeed one extends f across the boundary $\partial\Omega$ by imposing that the extended function F which agrees with f on Ω should be locally odd in the coordinate ρ which yields the signed distance to the boundary. Here locally odd means $F(y,\rho) = -F(y,-\rho)$ for ρ small enough. As the reader has guessed, (y,ρ) denotes local coordinates on a neighborhood of the boundary. The key fact which enters in the proof of Theorem 12 is that this odd extension operator is both continuous with respect to the Besov norm and the BV norm. An even extension operator would also be fine for the BV norm but certainly not for our Besov norm. Finally one applies Theorem 10 to this new function F, once it has been cut by a convenient cut-off function.

16 Wavelet coefficients of integrable functions

For a long time people believed in the following paradigm: if E is any functional Banach space for which the existence of an unconditional basis can be proved, then orthogonal wavelet bases provide effective unconditional bases. Then the fact that a given function belongs to E can be checked on size properties of its wavelet coefficients. Moreover a byproduct of this paradigm was that wavelet analysis was irrelevant for Banach spaces which do not possess unconditional bases. This belief was grounded by the history of the subject. Indeed Bernard Maurey proved the existence of an unconditional basis for the Hardy space $H^1(R)$ (in its real variable realisation) by abstract methods. Then Lennart Carleson discovered that an ad hoc wavelet like basis provided such an unconditional basis. A more systematic treatment was achieved by J.O. Strömberg who constructed the first orthonormal basis where the "mother wavelet" ψ_m was smooth (m continuous derivatives) with an exponential decay at infinity. In the same paper [16] Strömberg proved

that these bases were unconditional bases for $H^1(R)$. Strömberg construction extended to $H^1(R^n)$ and Strömberg discovered the rôle played by the scaling function φ. As it was already stressed the Banach space BV does not possess an unconditional basis and it was hard to believe that important results about wavelet coefficients of BV functions could ever be obtained. One can argue that BV has some intriguing similarities with the Hardy space $H^1(R^2)$. An even more surprising fact was discovered by Albert Cohen and Ronald DeVore. They proved that wavelet coefficients of $L^1(R^n)$ functions have some interesting properties. The normalization which will be used is the following. We write ψ_λ for $\psi(2^j x - k)$ and the wavelet coefficients of f are now $c(\lambda) =< f, \psi_\lambda >$. They are indexed by $\Lambda = Z \times Z^n \times F$ where F is a finite set with cardinality $2^n - 1$. Next we denote by Q_λ the corresponding dyadic cube defined by $\{x | 2^j x - k \in [0,1)^n\}$. The theorem on wavelet coefficients of $L^1(R^n)$ functions says the following:

Theorem 13 *For any real exponent γ larger than 1, there exists a constant $C_{\gamma,n}$ such that for f in $L^1(R^n)$ and for $\tau > 0$, one has*

$$\sum_{\{|c(\lambda)|>\tau |Q_\lambda|^\gamma\}} |Q_\lambda|^\gamma \leq C \frac{\|f\|_1}{\tau} \qquad (16.1).$$

In other words the wavelet coefficients $c(\lambda), \lambda \in \Lambda$, belong to to a weighted weak-l^1 space where the weighting factor is $|Q_\lambda|^\gamma$. Theorem 12 complements Theorem 11. Indeed Theorem 12 can be applied to the gradient of a BV function and the normalizations are adjusted in such a way that Theorem 13 corresponds to $\gamma > n$ in Theorem 11.

References

[1] Aline Bonami and S Poornima. Non-multipliers of Sobolev spaces. Journ. Funct. Anal. tome 71, (1987), 175-181.

[2] Christopher M. Brislawn. Fingerprints go digital. Notices of the AMS, Vol 42, no 11, Nov 1995, pp.1278-1283

web site http://www.c3. lanl. gov/ brislawn

[3] Albert Cohen. Numerical analysis of wavelet methods. Handbook of numerical analysis, P.G.Ciarlet and J.L.Lions eds. (1999)

[4] Albert Cohen and Robert Ryan. Wavelets and multiscale signal processing. Chapman and Hall, London (1995).

[5] Albert Cohen et al. Nonlinear approximation and the space BV. American Journal of Mathematics (to appear).

[6] Ingrid Daubechies. Ten lectures on wavelets. SIAM Philadelphia (1992).

[7] Ron DeVore. Nonlinear approximation. Acta Numerica (1998) pp. 51-150.

[8] Ron DeVore, Björn Jawerth and Vasil Popov. Compression of wavelet decompositions. American Journal of Mathematics 114 (1992) pp. 737-785

[9] Ron DeVore and Bradley J. Lucier. Fast wavelet techniques for near optimal image compression. 1992 IEEE Military Communications Conference October 11-14 (1992).

[10] David Donoho. Nonlinear Solution of Linear Inverse Problems by Wavelet-Vaguelette Decomposition, Applied and Computational Harmonic Analysis 2, 101-126 (1995).

[11] David Donoho, Ron DeVore, Ingrid Daubechies and Martin Vetterli. Data compression and harmonic analysis, IEEE Trans. on Information Theory, Vol. 44, No 6, October 1998, 2435-2476.

[12] David Donoho and Iain Johnstone. Wavelet shrinkage: Asymptopia? J.R.Statist. Soc. B (1995) 57 N 2 pp. 301-369.

[13] Jean-Pierre Kahane and Pierre-Gilles Lemarié-Rieusset. Fourier Series and Wavelets. Gordon and Breach Science Publishers (1996)

[14] Stéphane Mallat. A Wavelet Tour of Signal Processing. Academic Press (1998)

[15] Yves Meyer. Wavelets, Algorithms & Applications. SIAM, Philadelphia, (1993)

[16] Yves Meyer. Wavelets and Operators. Cambridge studies in advanced mathematics 37 CUP (1992)

[17] Yves Meyer. Wavelets, vibrations and scalings, CRM Monograph Series, Vol.9 (1998)

[18] Jean-Michel Morel. Filtres itératifs des images et quations aux drives partielles. Notes de cours du centre Emile Borel (1998)

[19] Jean-Michel Morel and Sergio Solimini. Variational methods in image segmentation. Birkhuser, Boston (1995).

[20] David Mumford. Book reviews on [16], Bulletin of the American Mathematical Society, Vol. 33, n2, April 1996.

[21] David Mumford and Jayant Shah. Boundary detection by minimizing functionals. Proc. IEEE Conf. Comp. Vis. Pattern Recognition, 1985.

[22] Frédéric Oru. Le rôle des oscillations dans quelques problèmes d'analyse non-linéaire. Thèse. CMLA. ENS-Cachan (9 Juin 1998)

[23] Stan Osher and Leonid Rudin. Total variation based image restoration with free local constraints. In Proc. IEEE ICIP, vol I, pages 31-35, Austin (Texas) USA, Nov.1994.

Yves Meyer
CMLA
Ecole normale supérieure de Cachan
94235 CACHAN Cedex
France