Harmonic Analysis of Two Problems in Signal Quantization and Compression

C. Sinan Güntürk

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
By the Program in
Applied and Computational Mathematics

November 2000

Abstract

This thesis is composed of two inpendendent parts:

Part I is on one-bit quantization of bandlimited functions, i.e., functions on the real line with compactly supported Fourier transforms. In such a scheme, a given bandlimited function x taking values in $[0, 1]$ is represented for each sampling density λ, by a $\{0, 1\}$-sequence q_λ such that convolving this sequence with an appropriately chosen filter kernel produces an approximation of the original function $\tilde{x}_\lambda$ which, as $\lambda \to \infty$, is required to converge to x in a given functional sense. A popular example of such a scheme is sigma-delta quantization, in which representative bitstreams are produced via a symbolic dynamics associated with a nonlinear discrete dynamical system forced by the input sample sequences. This thesis presents a new framework and improved techniques for the error analysis of sigma-delta quantization. A combination of tools from analytic number theory, harmonic analysis and dynamical systems are used to sharpen the existing error estimates.

Part II is on the functional space approach in the mathematical study of image compression. This approach, inspired by various toy models for natural images and the characterizations of linear and nonlinear approximation in wavelet bases through norm equivalences, treats images as functions in suitable Besov spaces. This thesis analyzes the validity and accuracy of this approach to a further extent, and demonstrates that while this is in general a fruitful approach, it can fail or be misleading in a variety of cases.

Acknowledgements

First, I would like to thank my advisor Ingrid Daubechies for her endless support during my entire study at Princeton. I cannot imagine another person with whom I could interact as much and as fruitfully as I did with her. She is truly a unique academic mentor in every possible sense.

Ron DeVore has been a constant source of inspiration for me. With his encouragement and wisdom, he taught me immeasurable amounts of beautiful mathematics in his wonderfully friendly style.

I interacted and collaborated with many researchers to whom I am indebted. I particularly thank Jeff Lagarias, Thao Nguyen and Vinay Vaishampayan for our fruitful collaboration, and my friend and colleague Özgür Yılmaz for his important share in our joint efforts in exploring a new problem. I also thank Zoran Cvetković, Jont Allen, Bob Adams, and Jade Vinson for various interesting discussions.

I thank Sergei Konyagin for showing me his result described in §4.2.3, and Wilhelm Schlag for suggesting to me a trick that simplified the proof of Theorem 3.11 and generalized the class of filters that could be used. I would also like to thank Peter Sarnak for his support and suggestions.

I was supported by various sources throughout my study; I particularly thank Princeton University for its 4-year long Francis Robbins Upton Fellowship. I thank AT&T Research and Robert Calderbank for the summer internship during which part of this work was done. I also thank the NSF KDI grant DMS-9872890, and the Princeton University Dean's Fund and APGA travel grant through which I was partially supported for academic travels.

Hakan Çağlar introduced me to research with much encouragement in the early years of my undergraduate study. He also suggested and motivated me to apply to Princeton. I thank him deeply.

During four years I spent in Fine Hall, my officemates have been some of my closest companions. I thank Ömer Alış, Igor Guskov, Jorge Silva and Matt DeVos for all the fun during the time we shared our offices.

I also met many nice people from PACM and Math Department. I particularly thank Toufic Suidan and Clíona Golden for their friendship. I also thank Tessy, Urmila, Jaime, Mike, Scott, Konstantinos, Radu, Fabrice, Jonathan, Jade and Slava for their pleasant company.

My housemates Refet and Taragay were a source of fun when I left Fine Hall, and our friends Özgür and İpek were frequently added. I thank them all. I also thank my friends Aykut, Hasan, Selçuk and Fatih for their close friendship over thousands of miles.

Finally, I thank my family for all their support, and especially my brother Serkan for sharing with me his interest in science from the very old days of our education.

To my family.

Contents

Part I

One-Bit Quantization of Bandlimited Signals

Chapter 1

Introduction

The first part of this thesis is on the approximation theory of oversampled quantization methods for bandlimited functions, with a particular emphasis on *sigma-delta* schemes. Our treatment of this subject will employ ideas and tools from a number of areas in pure and applied mathematics, ranging from harmonic analysis and approximation theory to dynamical systems, analytic number theory and information theory. Our aim is to give an integrated account of the various formulations of the problem as well as to separate and isolate some of its abstract ingredients, which turn out to lead to very interesting problems that can be studied on their own. We shall present many improvements to the previously known estimates of the performance of existing as well as new schemes.

***

Audio signals are modeled as bandlimited functions, i.e., real valued functions on the real line with compactly supported Fourier transforms. More formally, we define the class $\mathcal{B}_\Omega$ of bandlimited functions by

$$\mathcal{B}_\Omega = \{x : \mathbb{R} \to \mathbb{R} \mid \hat{x} \text{ is a finite Borel measure supported on } [-\Omega, \Omega]\}\,. \tag{1.1}$$

Here, $\hat{x}$ denotes the Fourier transform of x. A bandlimited function is completely determined by its sample values on a sufficiently dense uniform grid: Any $x \in \mathcal{B}_\Omega$ can be recovered from its samples $\{x(\frac{n}{\lambda})\}_{n\in\mathbb{Z}}$ for $\lambda > \Omega/\pi$, as a weighted sum of translates of a single kernel:

$$x(t) = \sum_{n\in\mathbb{Z}} x(\tfrac{n}{\lambda})\varphi(t - \tfrac{n}{\lambda}), \tag{1.2}$$

where φ is an appropriate function in $L^1(\mathbb{R})$. This is the classical *sampling theorem* whose precise statement we shall give in the next chapter.

Sampling is the first step of analog to digital conversion. The next step is *quantization*, which is the reduction of the sample values from their continuous range to a discrete set. The final step is *coding*, after which a digital bitstream is generated as the final representation of the continuous time signal. Quantization sometimes contains some of the ingredients of coding as well, as we shall see in the case of one-bit quantization.

In the simplest setting, the discrete set of quantization levels is an arithmetic progression $\mathcal{A}_\delta$ with spacing δ. Then, the approximate signal takes the form

$$\tilde{x}_{\delta,\lambda}(t) = \sum_{n\in\mathbb{Z}} q_{\delta,\lambda}(n)\varphi(t - \tfrac{n}{\lambda}), \tag{1.3}$$

where each $q_{\delta,\lambda}(n)$ is selected from the set $\mathcal{A}_\delta$. The selection procedure is usually "memoryless", meaning that there is a mapping $Q_\delta : \mathbb{R} \to \mathcal{A}_\delta$ for "rounding" such that $q_{\delta,\lambda}(n) = Q_\delta(x(\frac{n}{\lambda}))$. Classically, the trade-off between the size of the representation and the error of the approximation is controlled by the quantization step, while the sampling rate λ is fixed at a value λ_0 slightly above the critical Nyquist density Ω/π. Closer approximations are then obtained by letting $\delta \to 0$, to result in

$$\lim_{\delta\to 0} \tilde{x}_{\delta,\lambda_0} = x \tag{1.4}$$

in some functional sense.

The situation is reversed in an oversampled quantization scheme in which the quantization procedure is fixed and is coarse, and approximations are improved by increasing the sampling density only. For instance, $\mathcal{A}_\delta$ is fixed at say $\delta = \delta_0$, and the $q_{\delta_0,\lambda}(n)$ are appropriately chosen so that instead one now aims at achieving

$$\lim_{\lambda\to\infty} \tilde{x}_{\delta_0,\lambda} = x. \tag{1.5}$$

In the case when there is an a priori bound M for the maximum amplitude of the signals in consideration, the set $\mathcal{A}_{\delta_0}$ can be truncated to consist of the two values $\{-M, M\}$ only. This is what we shall assume from here on. Thus, we shall be interested only in *one-bit* schemes. It is by no means immediate that it is possible to construct such schemes. Indeed, the quantization procedure needs to be designed carefully; for instance, it can easily be checked that the simplest procedure of rounding every sample to the closest value in $\mathcal{A}_{\delta_0}$ is not capable of producing arbitrarily fine approximations to bandlimited functions as $\lambda \to \infty$. Schemes used in practice, called *sigma-delta modulators (quantizers)* are smart ways of circumventing this problem.

A sigma-delta modulator runs a nonlinear discrete dynamical system forced by the sample sequence $\{x(\frac{n}{\lambda})\}_{n\in\mathbb{Z}}$ of the function; an associated symbolic dynamics (the quantization) produces the bitstream of representation. The dynamical system typically takes the form of a difference equation

$$\Delta^k u(n) + q_{\delta_0,\lambda}(n) = x(\tfrac{n}{\lambda}), \tag{1.6}$$

where Δ is the difference operator defined by $\Delta u(n) = u(n) - u(n-1)$, and

$$q_{\delta_0,\lambda}(n) = Q(u(n-1), u(n-2), \ldots, x(\tfrac{n}{\lambda}), x(\tfrac{n-1}{\lambda}), \ldots) \tag{1.7}$$

for some appropriate quantizer function $Q(\cdot)$. There are various sigma-delta modulators of different orders and different rules of quantization given by k and $Q(\cdot)$, respectively; in some cases the $\Delta^k u(n)$ term is replaced by $(a * \Delta^k u)(n)$, where a is a fixed finite sequence and $*$ stands for the convolution of sequences. A common

crucial property of all the schemes is that an approximation to the original function is obtained via a convolutional operator acting on the quantized representative sequence, as given in (1.3). Although the dynamical systems are defined to fit such a reconstruction procedure, very little had been established, until recently, about the mathematical approximation theory for these schemes.

A sigma-delta scheme is called *stable* when the associated sequence u is bounded. Certainly, stability depends crucially on the function $Q(\cdot)$. Major problems start to arise for orders greater than one. Even for second order schemes, proving stability is a non-trivial task for most of the popular quantization rules used in practice. For orders of three and higher, ad hoc schemes have been designed in practice, but without any proof of stability [1, 2]. The first construction of a family of arbitrary order stable sigma-delta schemes is due to Daubechies and DeVore [3]; they also gave the first rigorous proof that the error of approximation for a stable k-th order sigma-delta scheme is bounded by $O(\lambda^{-k})$ in the L^∞ norm. We shall briefly describe their scheme in Chapter 4.

In the special case of first order schemes, the approximation estimate of [3] reduces to $O(\lambda^{-1})$. This falls short of the experimentally observed decay rate, which is $O(\lambda^{-3/2})$. (This estimate has always been stated in the folklore in an "averaged" sense.) One of the main results we shall present in this thesis is an improvement of the rigorous $O(\lambda^{-1})$ estimate towards this empirical decay rate. By employing various ideas and techniques from analytic number theory and harmonic analysis, we improve the pointwise estimate by raising the exponent from 1 to 4/3. This is the best result so far for first order schemes and for arbitrary bandlimited functions. We shall give a proof of this result in Chapter 3. For second order schemes, which have better approximation potential, more variety in the rules of quantization is possible. We fix on two particular choices, and analyze the corresponding dynamical systems for constant inputs. By applying techniques analogous to those for the first order scheme, we improve the error estimate both in the pointwise and the mean square sense.

One feature of the sigma-delta schemes is that the bits in the representation sequences play a "democratic" role in the convolutional reconstruction procedure. We analyze this aspect of the problem in an information theoretical framework in Chapter 4. A related problem in this context is the robustness of the symbolic dynamics arising from sigma-delta modulation under certain perturbations and with respect to the convolutional reconstruction procedure (1.3). This is of great practical importance as well as of theoretical interest. We provide tight upper and lower bounds on the mean square error of the optimal reconstruction for a uniformly distributed random constant input, and we prove its robustness under certain small systematic perturbations of the dynamical system. This result was obtained in collaboration with J. C. Lagarias and V. Vaishampayan [4]. This is a first step towards a theory of robust quantization, which will be complementary to the well-understood classical theory. Another interesting problem in sigma-delta quantization is connected to the approximation in a general L^p-norm; this will also be explored in Chapter 4.

Chapter 2

One-Bit Quantization: A General View

In the Introduction, we stated the problem of one-bit quantization in terms of general bandlimited functions. The problem remains interesting and still far from trivial when one considers only constant functions, which are, of course, an extreme special case of bandlimited functions. Indeed, a whole new set of problems, most of which strongly connect to number theory, arise in the study of this special case. Insight in these problems may shed more light on the general case, and certainly provides limits on what can be expected for more general bandlimited functions. In this chapter, we shall give a general account of one-bit quantization for both constant and arbitrary bandlimited functions.

2.1 On a Representation of Real Numbers

Consider the problem of representing real numbers in $[0, 1]$ by binary sequences in the following translation invariant manner: Each $x \in [0, 1]$ is mapped to a sequence $q := q_x \in \{0, 1\}^{\mathbb{Z}}$ such that for some appropriate sequence $h \in \ell^1(\mathbb{Z})$, called the *reconstruction filter*, one has

$$q * h = x, \tag{2.1}$$

where $*$ denotes convolution of two sequences, and the symbol x also denotes the constant sequence $(\ldots, x, x, \ldots)$. A natural normalization for h is the condition $\sum h(n) = 1$, which means that for each x, the "density" of 1's in the corresponding sequence q has to equal x. This also implies that the number 1 is necessarily represented by the sequence $(\ldots, 1, 1, \ldots)$ and the number 0 is necessarily represented by the sequence $(\ldots, 0, 0, \ldots)$. While these two sequences are the unique representations of these two numbers, there can be many possibilities for other values of x. For instance, the number $1/2$ may be represented by $(\ldots, 0, 1, 0, 1, \ldots)$, or by $(\ldots, 0, 0, 1, 1, 0, 0, 1, 1, \ldots)$, since for both cases, there are appropriate (and in fact, plenty of) choices for the reconstruction filter h so that (2.1) is satisfed.

On the other hand, as we shall show in §2.1.1, this problem is too strict in terms of the reconstruction formula (2.1) to be solvable for all x; we will see that a solution

exists if and only if x is rational, and the solution q is necessarily a periodic sequence. An alternative approach is to ask for a sequence of filters $(h_\lambda)_{\lambda>0}$ such that

$$q * h_\lambda \to x, \tag{2.2}$$

uniformly, or at least pointwise, as $\lambda \to \infty$. The normalization condition can be relaxed to the weaker form $\sum h_\lambda(n) \to 1$ as $\lambda \to \infty$. Clearly, there would not be any gain in introducing this alternative for a choice of sequence (h_λ) that converges in ℓ^1, since the problem would then be immediately reduced to the case (2.1). Indeed, a typical choice is $h_\lambda(n) = \frac{1}{\lambda}\chi_{[0,1)}(\frac{n}{\lambda})$, which converges to 0 uniformly as $\lambda \to \infty$, but not in ℓ^1. (We shall call this filter the *rectangular filter* or the *rectangular window* of length λ.) It turns out that in this new formulation, the problem has many solutions valid for *all* x; moreover, it is possible to employ "universal" filter sequences (h_λ) that remain the same for all values of x.

Yet another possibility is to let the binary representation vary with λ as well; that is, we require

$$q_\lambda * h_\lambda \to x \tag{2.3}$$

as $\lambda \to \infty$. We shall give examples of constructions in the case of this more general formulation later, when we discuss more general results in Chapter 4.

The last two settings are quite flexible and the main question consists in finding efficient representations in the sense that (2.2) or (2.3) converges rapidly in λ. We shall restrict ourselves to filters h_λ that are scaled versions of an averaging window, as in the example of rectangular filter we have just given. It is also desirable to determine the exact rate of convergence for particular schemes that have other features of interest.

2.1.1 General Considerations

Let us start with the problem (2.1). We claimed that a solution exists if and only if x is rational. This claim will follow simply as a corollary to a theorem of Szegö. We first recall the definition of spectral set for bounded sequences.

Definition 2.1. *Let a be a sequence in ℓ^∞. Then the spectral set $\sigma(a)$ is the set of all $\xi \in \mathbb{T}$ such that $b * a = 0$ ($b \in \ell^1$) implies $\sum_n b(n)e^{-in\xi} = 0$.*

The spectral set of a sequence a in ℓ^1 is precisely the closed support of its Fourier transform $\hat{a}(\xi) := \sum_n a(n)e^{-in\xi}$. Definition 2.1 extends this notion to arbitrary sequences in ℓ^∞, whose Fourier transforms are in general not functions. It is always true that $\sigma(a)$ is a closed subset of $\mathbb{T}$.

Szegö's theorem, as stated in the following form in [5, 6] by Helson, deals with spectral sets of sequences whose terms come from finite sets:

Theorem 2.2 (Szegö-Helson). *Let a be a sequence whose terms are all drawn from a finite set S of complex numbers. Unless the sequence is periodic, its spectral set fills $\mathbb{T}$.*

Let us apply this powerful theorem to the sequence $q - x$. Certainly, the terms of this sequence come from the finite set $\{-x, 1-x\}$. On the other hand, $h * x = x$, since $\sum h(n) = 1$. Thus, (2.1) is a restatement of $(q - x) * h = 0$. According to the conclusion of the theorem, a nonperiodic q (hence a nonperiodic $q - x$) would mean $\sigma(q - x) = \mathbb{T}$. This implies $\hat{h}(\xi) = \sum h(n)e^{-in\xi} = 0$ for all $\xi \in \mathbb{T}$, and hence $h \equiv 0$, a contradiction. Hence, only periodic solutions of (2.1) may exist.

Now, consider a periodic solution q, whose period is N. Then $q * h$ is also periodic and its period divides N. It is a simple calculation to show that

$$\sum_{n=1}^{N}(q * h)(n) = \sum_{n=1}^{N} q(n). \tag{2.4}$$

Since the left hand side is equal to Nx and the right hand side is an integer M such that $0 \leq M \leq N$, it follows that $x = M/N$, i.e., x is rational. This proves the "only if" part of our assertion. The "if" part follows from a trivial construction. Consider a rational number $x = M/N$ and let h be the rectangular averaging window of length N. Let q be a periodic sequence in $\{0, 1\}^{\mathbb{Z}}$ with period N and set exactly M elements of the set $\{q(k) : 1 \leq k \leq N\}$ to 1. Then, it is clear that $q * h = x$. In summary, we have proved the following:

Theorem 2.3. *There is a solution to (2.1) if and only if x is rational. The solution q is necessarily periodic and its period is a multiple of the denominator of x in its reduced form.*

The rectangular filter of length N is an admissible reconstruction filter for all the rational numbers in the set $\{P/Q : 0 \leq P \leq Q,$ and Q divides $N\}$. So, all the numbers in the Farey series[1] $\mathcal{F}_N$ can be decoded using a rectangular filter of length l.c.m.$(1, \ldots, N)$. Now, let us consider the implications of (2.4) on the filter h, that is, we would like to know under which conditions on the filter h, the formula (2.4) is satisfied for a given x. If a rational number $x = r/s$ (in its reduced form) is represented by a sequence q of period N, then s must divide N and precisely $M = Nr/s$ elements of the set $\{q(k) : k = 1, \ldots, N\}$ must be equal to 1. Let the corresponding indices be $l_1, \ldots, l_M$ and define $P(\xi) = \sum_{j=1}^{M} e^{-il_j\xi}$. Then, a straightforward calculation shows that $q * h = x$ implies $P(\xi)\hat{h}(\xi) = 0$ for all $\xi = 2\pi p/N$, $p = 1, \ldots, N-1$. If $M \neq N$ (i.e., $x \neq 1$), then $\hat{h}(\xi)$ must vanish at least at one of these points. Indeed, for almost any choice of $\{l_1, \ldots, l_M\}$, almost all of these roots would have to belong to $\hat{h}$. Since $\hat{h}$ can not admit a dense set of zeros in $\mathbb{T}$ when $h \in \ell^1$, this leads us to conjecture the following:

Conjecture 2.4. *It is impossible to construct a "universal" filter that can decode all rationals simultaneously. That is, there is no collection of sequences in $\{0, 1\}^{\mathbb{Z}}$ such that for every $x \in \mathbb{Q} \cap [0, 1]$, there is a sequence q in this collection for which $q * h = x$, where h is fixed.*

[1]The Farey series $\mathcal{F}_N$ of order N is the ascending series of irreducible fractions between 0 and 1 whose denominators do not exceed N [7].

Remark: One may weaken this conjecture by considering only filters $h \in \ell^1$ with exponential decay (which would require $\hat{h}$ to be analytic). On the other hand, an independent (trivial) argument easily rules out all finitely supported filters.

We now consider the alternative formulation (2.2). We have already stated in the introduction that in this more flexible approach it is possible to find solutions for all $x \in [0,1]$; even stronger, universal filter sequences (h_λ) can be employed. Before looking at sigma-delta quantization more closely as a scheme that generates such solutions, let us state the following result, which is merely an extension of Theorem 2.3.

Theorem 2.5. *A solution q to (2.2) for an irrational x is necessarily non-periodic.*

Proof. Suppose (2.2) is satisfied for some x and a periodic $q \in \{0,1\}^{\mathbb{Z}}$. Let N be the period of q. Then, similar to (2.4), we have

$$\sum_{n=1}^{N} (q * h_\lambda)(n) \to \sum_{n=1}^{N} q(n), \quad \text{as } \lambda \to \infty. \tag{2.5}$$

Combined with (2.2), this implies $x = \frac{1}{N} \sum_{n=1}^{N} q(n)$, i.e. $x \in \mathbb{Q}$. □

2.1.2 Sigma-Delta ($\Sigma\Delta$) Quantization

We briefly described a typical $\Sigma\Delta$ scheme in the Introduction. In this chapter, we shall be interested only in the basic properties of the first order $\Sigma\Delta$ scheme, which given a sequence $(x(n))_{n\in\mathbb{Z}}$ taking values in $[0,1]$, constructs a binary sequence q such that

$$\sum_{n_1}^{n_2} x(n) \sim \sum_{n_1}^{n_2} q(n) \tag{2.6}$$

for all n_1 and n_2. Here, we have used the symbol $\sim$ to mean that the two running sums differ from each other at most by a fixed amount that is uniform for all n_1 and n_2. This is done by the following procedure: Define the sequences X, Q and q by

$$X(n) := \sum_{m=1}^{n} x(m), \tag{2.7}$$

$$Q(n) := \lfloor X(n) \rfloor, \quad \text{and} \tag{2.8}$$

$$q(n) := Q(n) - Q(n-1). \tag{2.9}$$

Since x takes values in $[0,1]$, we have $q(n) \in \{0,1\}$; and at the same time (2.6) is satisfied, up to an error less than 1. X can be defined naturally for negative indices as well, by integrating backwards. Note that the equations (2.7), (2.8) and (2.9) correspond to "Σ", "quantization" and "Δ", respectively; hence giving the name of the scheme.

Define the auxiliary variable $u(n) := X(n) - Q(n)$. From (2.8), $u(n)$ is equal to the fractional part of $X(n)$, which we denote by $\langle X(n) \rangle$. In practice, neither $X(n)$

nor $Q(n)$ are computed in an electronic circuit, since these variables are in general unbounded. However, the sequence u is bounded and satisfies the recursion relation

$$u(n) - u(n-1) = x(n) - q(n), \qquad u(0) = 0. \tag{2.10}$$

In fact, this recursion is taken as the starting point in practice. One asks for a bounded solution u of (2.10) such that $q \in \{0,1\}^{\mathbb{Z}}$. The particular construction of q we have considered in (2.7)-(2.9) is just one of the solutions of (2.10). This solution can also be constructed by requiring $u(n)$ to satisy (2.10) and to lie in the interval $[0,1)$ for all n. Then, $q(n)$ is automatically given by

$$q(n) = \begin{cases} 1 & \text{if } u(n-1) + x(n) \geq 1, \\ 0 & \text{if } u(n-1) + x(n) < 1. \end{cases} \tag{2.11}$$

We shall now restrict the $\Sigma\Delta$ algorithm to the case in which x is a constant sequence. Our setting is (2.2), i.e., we shall reconstruct each $x \in [0,1]$ from its representative binary sequence q as the limit $\lim_{\lambda\to\infty} q * h_\lambda$ where the family (h_λ) of filters obey the scaling relation $h_\lambda(n) := \frac{1}{\lambda}\varphi(\frac{n}{\lambda})$ for an appropriate function $\varphi \in L^1$.

For a given filter h_λ, let us derive an estimate for the error $e_\lambda := x - q * h_\lambda$. The error may be bounded by a sum of two contributions:

$$|e_\lambda(n)| \leq |x(1 - \sum_n h_\lambda(n))| + \Big| \sum_k (x - q(k)) h_\lambda(n-k) \Big|. \tag{2.12}$$

Let us call these two terms e^1_λ and e^2_λ. It is possible to choose φ such that the first error term is zero for all λ. For instance, $\varphi = \chi_{[0,1]}$ has this property for all integer λ; on the other hand, $\varphi \in BV \cap \mathcal{B}_\pi$ with $\hat{\varphi}(0) = 1$ has it for all real $\lambda > 1$ (which may easily be seen using Poisson's summation formula). Assume a choice of φ with this property, which leaves us with $e^2_\lambda = (x - q) * h_\lambda$.

Theorem 2.6. *For all λ, $\|e_\lambda\|_{\ell^\infty} \leq \frac{1}{\lambda}\mathrm{Var}(\varphi)$.*

Proof. Let Δ denote the difference operator acting on sequences, defined by $\Delta u(n) := u(n) - u(n-1)$. Using the recursion relation (2.10), and that u is bounded by 1, it follows that

$$\begin{aligned} \|e_\lambda\|_{\ell^\infty} &= \|\Delta u * h_\lambda\|_{\ell^\infty}, \\ &= \|u * \Delta h_\lambda\|_{\ell^\infty}, \qquad (2.13) \\ &\leq \|u\|_{\ell^\infty} \|\Delta h_\lambda\|_{\ell^1}, \\ &\leq \frac{1}{\lambda}\mathrm{Var}(\varphi). \qquad (2.14) \end{aligned}$$

□

We call this the "basic estimate", in the sense that only boundedness of u was used in the derivation. In the next chapter, we shall improve the exponent of λ by examining the expression (2.13) more closely.

It is natural to ask how much information about x is "stored" in the first N bits of the binary sequence q, instead of checking the error of reconstruction for the specific convolution formulas considered above. So, let us consider the bit sequence $(q(1), \ldots, q(N))$. It is clear that the integer-valued N-tuple $(Q(1), \ldots, Q(N))$ contains the same amount of information as the N-tuple $(q(1), \ldots, q(N))$, since each can be recovered from the other. It is also clear that

$$\mathbf{Q}_x(N) := (Q(1), \ldots, Q(N)) = (\lfloor x \rfloor, \ldots, \lfloor Nx \rfloor) \tag{2.15}$$

is a monotonic function of x in the sense that if $x_1 \geq x_2$, then $\mathbf{Q}_{x_1}(N) \geq \mathbf{Q}_{x_2}(N)$ (meaning the inequality holds for all coordinates). It is easy to check that the number of distinct vectors in the set $\{\mathbf{Q}_x(N) : 0 \leq x \leq 1\}$ is exactly given by the number of distinct couples (n, r) such that there is a value of $x \in [0, 1]$ for which $\lfloor nx \rfloor = r$, where n ranges between 1 and N. This number is the same as the number of lattice points inside the triangle defined by $\{(n, r) : 1 \leq n \leq N;\, 0 \leq r \leq n\}$, which is equal to $N(N+1)/2$. Thus, there are only $O(N^2)$ possible distinct values for $\mathbf{Q}_x(N)$. In fact, each of these values correspond to an interval in $[0, 1]$ defined by the Farey series $\mathcal{F}_N$. This is the *scalar quantizer*[2] corresponding to the first N bits produced by a first order $\Sigma\Delta$ scheme.

It is well known that the lengths of intervals produced by the Farey series vary between $1/N^2$ and $1/N$. This means that for each N, the smallest possible error interval cannot be smaller than $1/N^2$: even for the most favorable $(q(1), \ldots, q(N))$, the value of x giving rise to these $q(k)$ cannot be determined with accuracy better than $O(1/N^2)$. In the next chapter, when we improve the basic estimate for the error of convolutional reconstruction, we shall see that this lower bound can in fact be achieved (up to a logarithmic factor) as $N \to \infty$.

2.2 Arbitrary Bandlimited Functions

2.2.1 General Setup

The general formulations given by (2.2) and (2.3) lead to an extension of the problem of one-bit quantization to more general functions than constants. Let $x(\cdot)$ be a function on $\mathbb{R}$, taking values in $[0, 1]$. Given a sequence u, we define the measure $\mu_\lambda(u)$ by

$$\mu_\lambda(u) := \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} u(n) \delta_{n/\lambda}, \tag{2.16}$$

where δ_a denotes the Dirac mass at the point a. Then, for each function x in some appropriate class $\mathcal{C}$, the problem is to find a family (q_λ) of binary representations such that, for a fixed, pre-chosen filter φ in $L^1(\mathbb{R})$, we have

$$(\mu_\lambda(q_\lambda) * \varphi)(\cdot) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_\lambda(n) \varphi(\cdot - \tfrac{n}{\lambda}) \quad \longrightarrow \quad x(\cdot) \tag{2.17}$$

[2] A scalar quantizer is a partition of an interval $[a, b]$ into subintervals and an associated sequence of representative points for each subinterval in this partition.

in a given functional sense, as $\lambda \to \infty$. (More generally, we can replace φ by a sequence (φ_λ) of filters.) Note that (2.3) is already contained in (2.17) if we define the discrete filter h_λ by $h_\lambda(n) = \frac{1}{\lambda}\varphi(\frac{n}{\lambda})$, and restrict our attention to constant functions.

Functions for which there exist solutions to the above problem include bandlimited functions, which was defined in Chapter 1 as

$$\mathcal{B}_\Omega = \{x : \mathbb{R} \to \mathbb{R} \mid \hat{x} \text{ is a finite Borel measure supported on } [-\Omega, \Omega]\}, \tag{2.18}$$

where $\hat{x}$ denotes the Fourier transform of x. Below, we list some of the important properties of bandlimited functions to which we shall frequently refer in this thesis.

1. A bandlimited function $x \in \mathcal{B}_\Omega$ is the restriction to $\mathbb{R}$ of an entire function of exponential type Ω. That is, the function defined by

$$x(z) = \frac{1}{2\pi}\int e^{-i\xi z} d\hat{x}(\xi), \quad z = x + iy, \tag{2.19}$$

   is entire and satisfies the growth bound $x(z) = O(e^{\Omega|y|})$.

2. The sampling theorem. In the literature, this theorem is more often stated for L^2 bandlimited functions, and not necessarily for the more general class $\mathcal{B}_\Omega$ defined above. Below, we provide a statement and proof of the sampling theorem in this more general setting. (In our discussion here and all the rest of this thesis, we will work with the space $\mathcal{B}_\pi$ to ease the notation. The analyses we present can always be transposed to arbitrary Ω by rescaling.)

   Theorem 2.7. *Let x be in $\mathcal{B}_\pi$, $\lambda > 1$, and φ a function in $L^1(\mathbb{R})$ such that $\hat{\varphi}$ satisfies*

$$\hat{\varphi}(\xi) = \begin{cases} 1 & \text{if } |\xi| \le \pi, \text{ and} \\ 0 & \text{if } |\xi| \ge \lambda\pi. \end{cases} \tag{2.20}$$

   Then, the following equality holds in the Cesàro mean for all t:

$$x(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} x(\tfrac{n}{\lambda})\varphi(t - \tfrac{n}{\lambda}), \tag{2.21}$$

   Proof. Let $\mathbf{B}$ be the space of continuous functions on $[-\lambda\pi, \lambda\pi]$ identified with the 1-torus. Its dual $\mathbf{B}^*$ is then the space of finite Borel measures on $[-\lambda\pi, \lambda\pi]$. $\mathbf{B}$ is a *homogeneous* Banach space on $[-\lambda\pi, \lambda\pi]$ (see [8] for a definition), so that for any $f \in \mathbf{B}$ and $\mu \in \mathbf{B}^*$, Parseval's theorem holds in the Cesàro mean [8, p. 35], i.e.,

$$\langle f, \mu\rangle = \lim_{N\to\infty} \sum_{n=-N}^{N} \Big(1 - \frac{|n|}{N+1}\Big)\hat{f}(n)\overline{\hat{\mu}(n)}\ , \tag{2.22}$$

   where $\hat{f}(n)$ and $\hat{\mu}(n)$ denote the n^{th} Fourier coefficient of f and μ, defined by $\hat{\mu}(n) := \overline{\langle e^{in\cdot/\lambda}, \mu\rangle}$, and $\langle f, \mu\rangle := \frac{1}{2\pi\lambda}\int_{-\lambda\pi}^{\lambda\pi} f\overline{d\mu}$. For each $t \in \mathbb{R}$, set $f(\xi) = \hat{\varphi}(\xi)e^{i\xi t}$ and $\mu = \overline{\hat{x}}$. Then $\langle f, \mu\rangle = \frac{1}{\lambda}x(t)$, $\hat{f}(n) = \frac{1}{\lambda}\varphi(t - \frac{n}{\lambda})$ and $\overline{\hat{\mu}(n)} = \frac{1}{\lambda}x(\frac{n}{\lambda})$, so that (2.21) (in the Cesàro mean) follows from (2.22). □

Remark: The formula (2.21) holds pointwise everywhere for the Cesàro mean, and hence for all values of t for which the right hand side converges. Typically, $\hat{\varphi}$ is chosen to be smooth so that the corresponding fast decay of φ enables an almost "local" reconstruction of x from the samples $\{x(\frac{n}{\lambda})\}_{n\in\mathbb{Z}}$, which removes any concern about the method of summation. (The formula (2.21) holds also in the L^2 sense when $x \in L^2(\mathbb{R})$, including the case $\lambda = 1$. However, in this critical sampling case, a smooth $\hat{\varphi}$ cannot be chosen; $\hat{\varphi} = \chi_{[-\pi,\pi]}$ is the only candidate.)

3. Bernstein's inequality [8]: If $f \in \mathcal{B}_\Omega$, then $\|f^{(s)}\|_{L^\infty} \leq \Omega^s \|f\|_{L^\infty}$. An L^p version also exists ([9, p. 14]).

We shall study solutions of (2.17) for the class

$$\mathcal{B}_\pi(\mathbb{R}, [0,1]) := \{x : \mathbb{R} \to [0,1] \mid x \in \mathcal{B}_\pi\}. \tag{2.23}$$

In contrast to the constant function case, all known methods for generating solutions for this more general class fall under $\Sigma\Delta$ schemes. (A brief definition for a general k-th order $\Sigma\Delta$ scheme was given in the Introduction.) Next, we give a basic analysis of the first order scheme. Improvements will be presented in Chapter 3.

2.2.2 $\Sigma\Delta$ Quantization: Basic Estimates

First Order

We described the first order $\Sigma\Delta$ algorithm in (2.7)-(2.9) for an input sequence x. The notation x will now refer to a function in $\mathcal{B}_\pi(\mathbb{R}, [0,1])$, and x_λ to the sequence of samples defined by $x_\lambda(n) = x(\frac{n}{\lambda})$. Similarly, we use the notations X_λ, Q_λ and q_λ to denote the quantities defined by

$$X_\lambda(n) := \sum_{m=1}^{n} x_\lambda(m), \tag{2.24}$$

$$Q_\lambda(n) := \lfloor X_\lambda(n) \rfloor, \quad \text{and} \tag{2.25}$$

$$q_\lambda(n) := Q_\lambda(n) - Q_\lambda(n-1). \tag{2.26}$$

Again, we assume that by integrating backwards, X_λ is defined for the negative indices as well.

In this section, we generalize the results of the previous section to arbitrary band-limited functions. The difference equation now reads as

$$u_\lambda(n) - u_\lambda(n-1) = x_\lambda(n) - q_\lambda(n), \qquad u(0) = 0; \tag{2.27}$$

where $u_\lambda(n)$ is defined to be $X_\lambda(n) - Q_\lambda(n) = \langle X_\lambda(n) \rangle$. Clearly, u_λ takes its values in $[0,1]$. Our setting is (2.17), and in general we will allow φ to depend on λ, which will then be denoted by φ_λ. We start with the corresponding "basic estimate" given in [3], where it is possible to use a fixed filter φ for all λ.

Theorem 2.8 ([3]). *Let $x \in \mathcal{B}_\pi$ with $0 \le x(t) \le 1$ for all t, and $\varphi \in BV(\mathbb{R})$ satisfying (2.20) for some fixed $\lambda_0 > 1$. Then for all $\lambda \ge \lambda_0$, one has*

$$\|x - \mu_\lambda(q_\lambda) * \varphi\|_{L^\infty} \le \frac{1}{\lambda}\mathrm{Var}(\varphi). \tag{2.28}$$

Proof. The sampling theorem states that $x = \mu_\lambda(x_\lambda) * \varphi$ for all $\lambda \ge \lambda_0$. Let Δ_η be the difference operator whose action on a measure is given by $\Delta_\eta \mu(\cdot) = \mu(\cdot) - \mu(\cdot - \eta)$, and let $\mathbf{1}$ denote the constant sequence of 1's. Then

$$\begin{aligned} x - \mu_\lambda(q_\lambda) * \varphi &= \mu_\lambda(x_\lambda - q_\lambda) * \varphi, \\ &= \mu_\lambda(\Delta u_\lambda) * \varphi, \\ &= \Delta_{1/\lambda}\mu_\lambda(u_\lambda) * \varphi, \\ &= \mu_\lambda(u_\lambda) * \Delta_{1/\lambda}\varphi, \end{aligned} \tag{2.29}$$

so that

$$\begin{aligned} \|x - \mu_\lambda(q_\lambda) * \varphi\|_{L^\infty} &\le \|\mu_\lambda(\mathbf{1}) * |\Delta_{1/\lambda}\varphi|\ \|_{L^\infty}, \\ &\le \frac{1}{\lambda}\mathrm{Var}(\varphi). \end{aligned} \tag{2.30}$$

In the last step, we made use of the identity $\mu_\lambda(\mathbf{1}) * f = \frac{1}{\lambda}\sum f(\cdot - \frac{n}{\lambda})$. □

Basic estimate for stable k-th order schemes

A stable k-th order $\Sigma\Delta$ scheme outputs a bit sequence q_λ that satisfies the difference equation

$$\Delta^k u_\lambda(n) = x_\lambda(n) - q_\lambda(n) \tag{2.31}$$

for a bounded sequence u_λ. The first construction of stable $\Sigma\Delta$ schemes for all orders is due to Daubechies and DeVore [3]. We shall return to their construction later in Chapter 4. The following is the corresponding basic estimate for a stable k-th order scheme:

Theorem 2.9 ([3]). *Let $x \in \mathcal{B}_\pi$ with $0 \le x(t) \le 1$ for all t, and suppose $\varphi \in BV(\mathbb{R})$ satisfies (2.20) for some fixed $\lambda_0 > 1$. Then, for any sequence q_λ that satisfies (2.31) for some sequence $u_\lambda \in \ell^\infty$, the following estimate holds:*

$$\|x - \mu_\lambda(q_\lambda) * \varphi\|_{L^\infty} \le \frac{1}{\lambda^k}\|u_\lambda\|_{\ell^\infty}\|\varphi^{(k)}\|_{L^1}\ . \tag{2.32}$$

Proof. The proof follows the same ideas as in the first order case. The identity in (2.29) now becomes

$$x - \mu_\lambda(q_\lambda) * \varphi = \mu_\lambda(u_\lambda) * \Delta^k_{1/\lambda}\varphi. \tag{2.33}$$

This implies

$$\|x - \mu_\lambda(q_\lambda) * \varphi\|_{L^\infty} \le \|u_\lambda\|_{\ell^\infty}\|\mu_\lambda(\mathbf{1}) * |\Delta^k_{1/\lambda}\varphi|\ \|_{L^\infty},$$

$$
\begin{aligned}
&\leq \frac{1}{\lambda}\|u_\lambda\|_{\ell^\infty}\mathrm{Var}(\Delta_{1/\lambda}^{k-1}\varphi),\\
&\leq \frac{1}{\lambda}\|u_\lambda\|_{\ell^\infty}\|\Delta_{1/\lambda}^{k-1}\varphi'\|_{L^1},\\
&\leq \frac{1}{\lambda^k}\|u_\lambda\|_{\ell^\infty}\|\varphi^{(k)}\|_{L^1}. \qquad (2.34)
\end{aligned}
$$

In the last step, we have made use of the fact that

$$
\begin{aligned}
\|\Delta_{1/\lambda}^{k-1}\varphi'\|_{L^1} &\leq \omega_{k-1}(\varphi', \lambda^{-1})_{L^1}\\
&\leq \lambda^{-(k-1)}|\varphi'|_{W_1^{k-1}}\ , \qquad (2.35)
\end{aligned}
$$

where $\omega_r(f,\cdot)_{L^p}$ denotes the *r-th modulus of smoothness of f in L^p* (see, [10, p. 44] for a definition) and the final step is due to the well-known inequality:

$$
\omega_r(f,t)_{L^p} \leq t^r|f|_{W_p^r}, \qquad (2.36)
$$

which can be found in [10, p. 46]. □

Chapter 3

Improving Error Estimates for Sigma-Delta Systems

This chapter presents some of the main results of the first part of this thesis; namely, improvements on the basic error estimates for $\Sigma\Delta$ systems that were given in Chapter 2. These improvements will heavily utilize the theory of uniform distribution for point sequences and stationary phase methods for exponential sums. Below, we summarize the basic definitions and theorems that we shall use in the proofs.

3.1 Preliminaries From The Theory of Uniform Distribution and Exponential Sums

Let $\{u_n\}_{n=1}^{\infty}$ be a sequence of points in $[0,1)$ identified with the 1-torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. The sequence $\{u_n\}$ is said to be *uniformly distributed* (in short, *u.d.*) if

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N \ : \ u_n \in I\}}{N} = |I| \tag{3.1}$$

for every arc I in $\mathbb{T}$. Define the N-term *discrepancy* of the sequence $\{u_n\}$ as

$$D_N := D_N(\{u_n\}) := \sup_{I\subset\mathbb{T}} \left| \frac{\#\{1 \le n \le N \ : \ u_n \in I\}}{N} - |I| \right|. \tag{3.2}$$

It is an elementary result that $\{u_n\}$ is u.d. if and only if $D_N(\{u_n\}) \to 0$ as $N \to \infty$. Equivalent characterizations of uniform distribution are given by *Weyl's criterion*:

Theorem 3.1 (Weyl).

$$\{u_n\} \text{ is u.d.} \iff \frac{1}{N}\sum_{n=1}^{N} e^{2\pi i k u_n} \to 0 \quad \text{for each nonzero } k \in \mathbb{Z}, \tag{3.3}$$

$$\iff \frac{1}{N}\sum_{n=1}^{N} f(u_n) \to \int_{\mathbb{T}} f(u)\,du \quad \text{for every Riemann-integrable (or, equivalently, continuous) } f \text{ on } \mathbb{T}. \tag{3.4}$$

These are "qualitative" statements. The relation between how good the distribution of a sequence is and how fast (3.3) and (3.4) converge are studied in the "quantitative" theory. The second Weyl criterion is especially relevant to numerical integration. The following two results are fundamental quantitative measures in the theory:

Theorem 3.2 (Koksma's inequality). *For any function $f \in BV([0,1])$ and a finite sequence of points $u_1, \ldots, u_N$ in $[0,1]$,*

$$\Big|\frac{1}{N}\sum_{n=1}^{N} f(u_n) - \int_0^1 f(u)du\Big| \leq Var(f) D_N, \tag{3.5}$$

where D_N denotes the discrepancy *of the sequence $u_1, \ldots, u_N$ and $Var(f)$ is the total variation of f.*

Theorem 3.3 (Erdős-Turán inequality). *The discrepancy D_N of any real numbers $u_1, \ldots, u_N$ is bounded by*

$$D_N \leq C \inf_{K \geq 1} \left(\frac{1}{K} + \sum_{k=1}^{K} \frac{1}{k} \Big| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i k u_n} \Big| \right), \tag{3.6}$$

for some absolute constant C.

Multidimensional discrepancy

The theory of uniform distribution generalizes naturally to higher dimensions, however with some added complexity. For our analysis of higher order sigma-delta schemes, we shall need the multidimensional versions of the theorems listed above. We will be working mostly in two dimensions.

Let $\{\mathbf{u_n}\}$ be a sequence in $[0,1)^d$ identified with $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. For a measurable subset H of $[0,1)^d$, define

$$D_N(H) := \left| \frac{\#\{1 \leq n \leq N \ : \ \mathbf{u}_n \in H\}}{N} - |H| \right|, \tag{3.7}$$

where $|H|$ denotes the d-dimensional Lebesgue measure of H. Let $\mathcal{I}^d$ denote the set of all intervals (i.e., the set of all rectangles whose sides are parallel to the axes) in $\mathbb{T}^d$. The discrepancy D_N is by definition

$$D_N = \sup_{H \in \mathcal{I}^d} D_N(H). \tag{3.8}$$

The sequence $\{\mathbf{u_n}\}$ is said to be u.d. if the condition $\lim_{N\to\infty} D_N(H) = 0$ holds for every $H \in \mathcal{I}^d$. Again, this is equivalent to $\lim_{N\to\infty} D_N = 0$. Weyl's criterion naturally extends using multidimensional versions of (3.3) and (3.4).

A definition of discrepancy exists also for arbitrary non-negative Borel measures μ on $[0,1)^d$. The discrepancy of μ with respect to the set $H \in [0,1)^d$, denoted by $D(\mu; H)$, is defined to be $|\mu(H) - |H||$. Similarly, one has the definition

$$D(\mu) := \sup_{H \in \mathcal{I}^d} D(\mu; H) \tag{3.9}$$

for the discrepancy of μ. Clearly, $D_N(\{\mathbf{u}_n\}) = D(\mu_N)$, where μ_N is defined by $\mu_N(A) := \frac{1}{N}\sum_{n=1}^{N} \chi_A(\mathbf{u}_n)$ for $A \subset \mathbb{T}^d$.

If the supremum in (3.9) is taken instead over all *convex* subsets of $\mathbb{T}^d$, then this quantity defines the *isotropic* discrepancy $J(\mu)$. Clearly, one has $D(\mu) \leq J(\mu)$; on the other hand, an inequality in the reverse direction exists only in a weaker sense: $J(\mu) \leq C_d D(\mu)^{1/d}$, where C_d is a constant that depends only on the dimension d.

While the family of convex sets is much larger than the family of intervals, we will need to be able to work with an even larger class of sets to prove our results in Section 3.3. The family of sets whose topological boundaries have zero Lebesgue measure (i.e. Jordan measurable sets) will suit our purposes. Every such set in this family belongs to a sub-family $\mathcal{M}_b$ of sets $H \subset [0,1)^d$ for which $|\{\mathbf{u} \in H^c : \mathrm{dist}(\mathbf{u}, H) < \epsilon\}| \leq b(\epsilon)$ and $|\{\mathbf{u} \in H : \mathrm{dist}(\mathbf{u}, H^c) < \epsilon\}| \leq b(\epsilon)$ for every $\epsilon > 0$, where $b : (0,\infty) \to (0,\infty)$ is a monotonically increasing function such that $\lim_{\epsilon \to 0^+} b(\epsilon) = 0$. The following theorem, found in [11, pp. 173] and also in [12], gives a discrepancy estimate for sets in such a family:

Theorem 3.4 (Niederreiter, Wills). *Let $b : (0,\infty) \to (0,\infty)$ be monotonically increasing such that $b(\epsilon) \geq \epsilon$ for all $\epsilon > 0$, and $\lim_{\epsilon \to 0^+} b(\epsilon) = 0$. Then, for every $H \in \mathcal{M}_b$, one has*

$$D(\mu; H) \leq 4\, b(2\sqrt{d} D(\mu)^{1/d}). \tag{3.10}$$

A multidimensional version of Koksma's inequality (called *the Koksma-Hlawka inequality*) holds for functions of *bounded variation in the sense of Hardy and Krause*. We will not go into the details but refer to [13, 14] only. (However, a "baby" version of this theorem will be employed later in Section 3.3.) On the other hand, a generalization of Erdős-Turán inequality is simpler to state and is given by the following:

Theorem 3.5 (Erdős-Turán-Koksma's inequality). *The discrepancy D_N of any real numbers $\mathbf{u}_1, \ldots, \mathbf{u}_N$ in $\mathbb{T}^d$ is bounded by*

$$D_N \leq C_d \inf_{K \geq 1} \left(\frac{1}{K} + \sum_{0 < \|\mathbf{k}\|_\infty \leq K} \frac{1}{r(\mathbf{k})} \Big| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i \mathbf{k} \cdot \mathbf{u}_n} \Big| \right) \tag{3.11}$$

for some absolute constant C_d, where $r(\mathbf{k}) := \prod_{i=1}^{d} \max\{1, |k_i|\}$ for $\mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{Z}^d$.

Exponential sums of a single variable: two theorems

It is clear from the Erdős-Turán inequality that the problem of estimating the discrepancy can be turned precisely into the problem of estimating certain exponential sums. In proving our results, we shall actually make use of only a relatively tiny section of the theory of exponential sums. With regards to sigma-delta quantization, two types of exponential sums will be relevant for us. The first type is the well-studied class of *Weyl sums*

$$S = \sum_{n=1}^{N} e^{2\pi i f(n)}, \tag{3.12}$$

where, by definition, f is a polynomial (with real coefficients). This type of sums will arise in $\Sigma\Delta$ schemes with constant input. The second type of sums are given by more general functions in the exponent, that are not necessarily polynomials, yet still have a certain amount of smoothness. These sums will arise when the input is an arbitrary bandlimited function. For both types of sums, extremely sophisticated tools are available in the mathematical literature to estimate their sizes. We shall require here only outcomes of more "general purpose" tools, for they already lead to substantial improvements of the basic estimates. We shall make use of the *truncated Poisson formula* and *van der Corput's Lemma*, which we give below; we shall have no need of more sophisticated tools, such as the *method of exponent pairs* (e.g. see [15]).

Theorem 3.6 (Truncated Poisson, [15]). *Let f be a real-valued function and suppose that f' is continuous and increasing on $[a, b]$. Put $\alpha = f'(a)$, $\beta = f'(b)$. Then*

$$\sum_{a \le m \le b} e^{2\pi i f(m)} = \sum_{\alpha - 1 \le \nu \le \beta + 1} \int_a^b e^{2\pi i (f(\tau) - \nu\tau)} d\tau + O(\log(2 + \beta - \alpha)). \tag{3.13}$$

(If f' is decreasing, taking the complex conjugate of the above expression applied to $-f$ leads to the same expression with α and β switched.)

Theorem 3.7 (van der Corput, [16]). *Suppose ϕ is real-valued and smooth in the interval (a, b), and that $|\phi^{(r)}(t)| \ge \mu$ for all $t \in (a, b)$ and for a positive integer r. If $r = 1$, suppose additionally that ϕ' is monotonic. Then there exists an absolute constant c_r such that*

$$\Big| \int_a^b e^{i\phi(t)}\, dt \Big| \le c_r \mu^{-1/r}. \tag{3.14}$$

Discrepancy of arithmetic progressions modulo 1:

Maybe the most important examples of uniformly distributed sequences are arithmetic progressions modulo 1, defined by $u_n = \langle n\alpha \rangle$, with $\alpha \in \mathbb{R} \backslash \mathbb{Q}$. These sequences

arise in the first order $\Sigma\Delta$ quantization with constant input and the corresponding discrepancy estimates will directly relate to the error estimates.

There will be two different types of statements for us: metric results which are valid for almost every α (with respect to the Lebesgue measure), and results that depend on the finer Diophantine properties of α. Below, we give a few basic definitions and important theorems that will be relevant for our purposes.

Denote by $\|u\|$ the metric $\mathrm{dist}(u, \mathbb{Z})$. Let $\psi : \mathbb{Z}^+ \to \mathbb{R}^+$ be a given function. The irrational number α is said to be *of type* $<\psi$ if the inequality $n\|n\alpha\| \geq 1/\psi(n)$ holds for all positive integers n. If α is of type $<\psi$ for a constant function ψ, then one says α is *of constant type*. A related concept is the following: Let $\eta \in \mathbb{R}^+ \cup \{\infty\}$. The number α is said to be *of type* η if $\eta = \sup\{\gamma : \liminf_{n\to\infty} n^\gamma \|n\alpha\| = 0\}$.

Let us use the notation $D_N(\alpha)$ for the N-term discrepancy of the sequence $\langle n\alpha\rangle$. It follows from the Erdős-Turán inequality that

$$D_N(\alpha) \leq C\left(\frac{1}{K} + \frac{1}{N}\sum_{k=1}^{K}\frac{1}{k\|k\alpha\|}\right), \tag{3.15}$$

for any positive integer K. The sum term $\sum_{k=1}^{K}\frac{1}{k\|k\alpha\|}$ can be estimated in terms of the type of α [13, Lemma 3.3]. Then it follows, for instance, that if α is of finite type η, then for every $\epsilon > 0$, the discrepancy satisfies $D_N(\alpha) = O(N^{-1/\eta+\epsilon})$ [13, pp. 123, Theorem 3.2]. This says that the sequence $\langle n\alpha\rangle$ shows a better distribution behaviour for α that are badly approximable by rationals. The celebrated theorem of Thue-Siegel-Roth[1] implies that every irrational algebraic number is of type $\eta = 1$, the smallest possible type attainable. Yet, better estimates are possible for irrationals of constant type (for instance, all quadratic irrationals). For these, the discrepancy satisfies $D_N(\alpha) = O(N^{-1}\log N)$. This is the smallest possible order of discrepancy for *any* infinite sequence $\{u_n\}$ due to the following lower bound: $D_N(\{u_n\}) \geq c\, N^{-1}\log N$ for infinitely many N, where c is an absolute constant.

An important metric result for us (due to Khinchine) is the following: Let $\epsilon > 0$ be given. Then, almost all α are of type $<C_\alpha \log^{1+\epsilon} 2q$, where C_α is a constant that may depend on α. This result may be used to prove the following theorem which will be used in the next section:

Theorem 3.8 ([13]). *For any $\epsilon > 0$, the discrepancy $D_N(\alpha) = O(N^{-1}\log^{2+\epsilon} N)$ for almost all α.*

Remark: It is easy to check that the same estimate holds uniformly (with the same constant) for any translate of the sequence $(\langle n\alpha\rangle)$. This strengthens the qualitative result that for irrational α, $(\langle n\alpha\rangle)$ is not only u.d. but also *well distributed* [13].

[1]For every irrational algebraic number α and for every $\epsilon > 0$, there exists a positive constant $c = c(\alpha, \epsilon)$ such that

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{c}{q^{2+\epsilon}}$$

for all integers $q > 0$ and p.

3.2 Improved Estimates for First Order Systems

Improving the Basic Estimate for Constants

The basic estimate for first order $\Sigma\Delta$ quantization with constant inputs was given in Theorem 2.6. Below is an improvement of this estimate:

Theorem 3.9. *Let $\epsilon > 0$ be given. Set $h_\lambda(n) = \frac{1}{\lambda}\varphi(\frac{n}{\lambda})$, where φ is the triangular filter defined by $\varphi(t) = (1 - |t|)\chi_{[-1,1]}(t)$. Then for almost every $x \in [0, 1]$, the error $e_\lambda = x - q * h_\lambda$ satisfies the estimate*

$$\|e_\lambda\|_{\ell^\infty} \leq C_x \lambda^{-2} \log^{2+\epsilon} \lambda. \tag{3.16}$$

Proof. We start with the error expression given in (2.13). First, note that for any constant c, one has $u * \Delta h_\lambda = (u - c) * \Delta h_\lambda$. For the triangular filter, and choosing $c = 1/2$, this expression can be rewritten as

$$((u - \frac{1}{2}) * \Delta h_\lambda)(n) = \frac{1}{\lambda^2} \sum_{k=0}^{\lambda-1} (u(n + k) - \frac{1}{2}) - \frac{1}{\lambda^2} \sum_{k=1}^{\lambda} (u(n - k) - \frac{1}{2}). \tag{3.17}$$

The choice $c = 1/2$ was made in order to exploit that the state variable $u(n) = \langle X(n)\rangle = \langle nx\rangle$ forms a uniformly distributed sequence in $[0, 1]$ for all irrational values of x and its average value is $1/2$. Koksma's inequality reduces the problem to considering the discrepancy values for the two sequences $u(n - \lambda), \ldots, u(n - 1)$ and $u(n), \ldots, u(n + \lambda - 1)$. The discrepancy can be bounded using the Erdős-Turán inequality, which gives

$$\Big|\frac{1}{\lambda} \sum_{k=a+1}^{a+\lambda} u(k) - \frac{1}{2}\Big| \leq \inf_K \; C \left(\frac{1}{K} + \sum_{k=1}^{K} \frac{1}{k} \Big| \frac{1}{\lambda} \sum_{m=1}^{\lambda} e^{2\pi i k m x} \Big| \right). \tag{3.18}$$

Note that the bound obtained in this way is uniform in a. This observation, together with (3.17) and Theorem 3.8 yields the desired result. □

A Mean Square Error Estimate for Constants

Let us use the notation $e_{\lambda,x}$ to denote the dependence of the error e_λ on the input value x. We assume that φ is the triangular filter, so that (3.17) and (3.18) hold. The precise behaviour of $e_{\lambda,x}$ can be described only by means of the continued fraction expansion of x (see, e.g. [13, 17, 14]). However, the mean behavior is simpler. Let us consider the mean squared error (MSE)

$$\mathrm{MSE}(e_\lambda) := \int_0^1 \|e_{\lambda,x}\|^2_{\ell^\infty} dx. \tag{3.19}$$

A straightforward bound for $\mathrm{MSE}(e_\lambda)$ follows directly from (3.18):

Theorem 3.10. *MSE*$(e_\lambda) \le C\lambda^{-3}\log^2\lambda$, *for some absolute constant* C.

Proof. Let $P_{\lambda,k}(x)$ denote the trigonometric polynomial $\lambda^{-1}\sum_{m=1}^{\lambda} e^{2\pi ikmx}$. Note that $\|P_{\lambda,k}\|_{L^2([0,1])} = \lambda^{-1/2}$. Then,

$$\begin{aligned}
\mathrm{MSE}(e_\lambda) &\le \frac{C}{\lambda^2}\int_0^1 \inf_K \Big(\frac{1}{K} + \sum_{k=1}^{K} k^{-1}|P_{\lambda,k}(x)|\Big)^2 dx \\
&\le \frac{C'}{\lambda^2}\inf_K \int_0^1 \left(\frac{1}{K^2} + \Big(\sum_{k=1}^{K} k^{-1}|P_{\lambda,k}(x)|\Big)^2\right) dx \\
&\le \frac{C'}{\lambda^2}\inf_K \left(\frac{1}{K^2} + \sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{kl}\int_0^1 |P_{\lambda,k}(x)||P_{\lambda,l}(x)|dx\right) \\
&\le \frac{C'}{\lambda^2}\inf_K \left(\frac{1}{K^2} + \frac{1}{\lambda}\log^2 K\right),
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality in the last step. Finally, by choosing $K \sim \lambda^{1/2}$, we arrive at the desired bound. □

Remark: The exponent of λ in this estimate is optimal. Indeed, using number theoretical tools, it is shown in [4] that

$$C_1\lambda^{-3} \le \int_0^1 |e_{\lambda,x}(0)|^2 dx \le C_2\lambda^{-3}. \tag{3.20}$$

We will give a proof of this result in Chapter 4, where we'll discuss in more detail several related results. It is natural to conjecture that the quantity defined by

$$\mathrm{MSE}'(e_\lambda) := \left\| \int_0^1 |e_{\lambda,x}(\cdot)|^2 dx \right\|_{l^\infty} \tag{3.21}$$

satisfies a similar estimate. (Note that $\mathrm{MSE}'(e_\lambda) \le \mathrm{MSE}(e_\lambda)$.) On the other hand, it is shown in [18] that the quantity

$$\int_0^1 \frac{1}{2N+1}\sum_{n=-N}^{N} |e_{\lambda,x}(n)|^2 dx \tag{3.22}$$

(which is even smaller) also behaves as $O(\lambda^{-3})$ as $N \to \infty$.

Improving the Basic Estimate for Bandlimited Functions

We shall apply the ideas of the previous section to prove the following theorem, which is an improvement of the basic estimate that was given in Theorem 2.8:

Theorem 3.11. *For all* $\eta > 0$, *there exists a family* $\{\varphi_\lambda\}_{\lambda\ge 1}$ *of filters such that, for all* x *in Theorem 2.8, and all* t *for which* $x'(t)$ *does not vanish, we have*

$$|x(t) - (\mu_\lambda(q_\lambda) * \varphi_\lambda)(t)| \le C\lambda^{-4/3+\eta} \tag{3.23}$$

for some constant $C = C(\eta, x'(t))$.

Note that bandlimited functions are analytic, so that the derivative x' of a non-constant bandlimited function x has at most countably many zeros, with no accumulation point. It is also possible to carry out a higher order analysis at the zeros of x', but we shall not go into this here.

As is customary, we shall use the notations $C, C', C_1, C_2, ...$ for generic constants that may change value from one proof to another; constants of different values occurring in the same argument will be distinguished by different indices.

Proof. We divide the proof into a number of steps.
1. Fix t (for which $x'(t) \neq 0$). For each λ, let $N_\lambda = \lfloor \lambda t \rfloor$, and define the sequence U_λ by

$$U_\lambda(n) - U_\lambda(n-1) = u_\lambda(n) - \frac{1}{2}, \qquad U_\lambda(N_\lambda) = 0. \tag{3.24}$$

Let also $t_\lambda = N_\lambda/\lambda$ and $\delta_\lambda = t - t_\lambda$. Note that $|\delta_\lambda| \leq 1/\lambda$. Now, (2.29) can be written as

$$\begin{aligned} x(t) - (\mu_\lambda(q_\lambda) * \varphi_\lambda))(t) &= \frac{1}{\lambda} \sum_n \Delta U_\lambda(n) \Delta_{1/\lambda} \varphi(t - \tfrac{n}{\lambda}), \\ &= \frac{1}{\lambda} \sum_n U_\lambda(n) \Delta^2_{1/\lambda} \varphi(t - \tfrac{n}{\lambda}), \qquad (3.25) \\ &= \frac{1}{\lambda} \sum_n U_\lambda(N_\lambda + n) \Delta^2_{1/\lambda} \varphi(-\tfrac{n}{\lambda} + \delta_\lambda), \qquad (3.26) \end{aligned}$$

for any φ that decays sufficiently fast. Denote the error expression by $e_\lambda(t)$.
2. Our purpose is to find non-trivial bounds for $U_\lambda(N_\lambda + n)$ by accounting for the cancellations in

$$U_\lambda(N_\lambda + n) = \sum_{m=1}^{n} \Big(u_\lambda(N_\lambda + m) - \frac{1}{2} \Big), \tag{3.27}$$

where we have assumed $n > 0$, the other case being essentially the same. Note that the trivial bound is $|n|/2$. We shall prove the following estimate:

$$|U_\lambda(N_\lambda + n)| \leq C_1 \Big(\lambda^{2/3} + \frac{\lambda^{1/2}}{|x'(t)|^{1/2}} \Big), \tag{3.28}$$

for all $n \leq C_2 |x'(t)| \lambda$ and $\lambda > C_3 |x'(t)|^{-1}$.

The inequalities of Koksma and Erdős-Turán result in the bound

$$|U_\lambda(N_\lambda + n)| \leq \inf_K \; C \left(\frac{n}{K} + \sum_{k=1}^{K} \frac{1}{k} \Big| \sum_{m=1}^{n} e^{2\pi i k u_\lambda(N_\lambda + m)} \Big| \right), \tag{3.29}$$

which reduces our task to analyzing the behaviour of the exponential sums

$$S_{\lambda,k}(n) := \sum_{m=1}^{n} e^{2\pi i k X_\lambda(N_\lambda + m)}, \tag{3.30}$$

since $u_\lambda(n) = \langle X_\lambda(n)\rangle$.

3. We will use the stationary phase methods of van der Corput to estimate the exponential sums given in (3.30).

In our case, X_λ is initially defined only on the integers; however, (2.7) immediately yields an (analytic) interpolation of X_λ:

Lemma 3.12. *The sequence X_λ can be extended to an analytic function, which we shall denote by X_λ as well. Moreover, for $\lambda \geq C|x'(t)|^{-1}$, and all real τ in the range $0 \leq \tau \leq C_3|x'(t)|\lambda$, one has*

$$C_1\frac{|x'(t)|}{\lambda} \leq |X_\lambda''(N_\lambda + \tau)| \leq C_2\frac{|x'(t)|}{\lambda}, \tag{3.31}$$

where C, C_1, C_2, and C_3 are absolute numerical constants.

Proof. Using the Taylor expansion of x about the point t_λ, we proceed as follows:

$$\begin{aligned} X_\lambda(N_\lambda + m) &= X_\lambda(N_\lambda) + \sum_{l=1}^{m} x(t_\lambda + \tfrac{l}{\lambda}) \\ &= X_\lambda(N_\lambda) + \sum_{s=0}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} \sum_{l=1}^{m} l^s. \end{aligned} \tag{3.32}$$

Note that the sum $P_s(m) := \sum_{l=1}^{m} l^s$ can be written as a polynomial $\sum_{j=1}^{s+1} P_{s,j} m^j$ of degree $s+1$ in m.[2] Then,

$$X_\lambda(N_\lambda + m) = X_\lambda(N_\lambda) + \sum_{s=0}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} \sum_{j=1}^{s+1} P_{s,j} m^j \tag{3.33}$$

$$= X_\lambda(N_\lambda) + \sum_{j=1}^{\infty} m^j \sum_{s=j-1}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} P_{s,j}. \tag{3.34}$$

[2]This elementary result can easily be seen from the identity

$$\sum_{l=1}^{m} l^s = \int_0^m x^s dx + \sum_{l=1}^{m} \int_{-1}^{0} [l^s - (x+l)^s] dx,$$

by expanding $(x+l)^s$ in powers of l and evaluating the integrals. This results in the recursion relation

$$P_s(m) = \frac{m^s}{s+1} - \sum_{j=1}^{s} \frac{(-1)^j}{j+1} \binom{s}{j} P_{s-j}(m),$$

which proves that $P_s(m)$ is a polynomial in m of degree $s+1$, and further leads to a recursion relation in terms of the coefficients $P_{s,j}$:

$$P_{s,s+1} = \frac{1}{s+1}, \qquad P_{s,k} = \sum_{j=k-1}^{s-1} \frac{(-1)^{s-j+1}}{s-j+1} \binom{s}{j} P_{j,k}, \quad k = 1, \ldots, s.$$

It then follows by a straightforward inductive argument that $|P_{s,j}| \leq s!/j!$. The numbers $P_{s,j}$ have explicit representations in terms of the *Bernoulli numbers*. Note that $P_{s,s} = 1/2$ for all $s \geq 1$.

We use this last expression to define

$$X_\lambda(N_\lambda+\tau) = X_\lambda(N_\lambda) + \sum_{j=1}^{\infty} a_j \tau^j \tag{3.35}$$

for all $\tau \geq 0$, where a_j is the sum term appearing in (3.34). The simple bound $|P_{s,j}| \leq s!/j!$ and Bernstein's inequality easily yield

$$|a_j| < \frac{2}{j!}\Big(\frac{\pi}{\lambda}\Big)^{j-1} \tag{3.36}$$

for $\lambda > 2\pi$. Let us show that

$$X'_\lambda(N_\lambda+\tau) = x(t_\lambda + \tfrac{\tau}{\lambda}) + R_\lambda(\tau), \tag{3.37}$$

where $R_\lambda(\tau)$ is small compared to $x(t_\lambda+\frac{\tau}{\lambda})$ for $\tau = O(\lambda)$. We start with noting that $P_{s,s+1} = 1/(s+1)$ for all s. Then, starting from (3.32),

$$\begin{aligned} X'_\lambda(N_\lambda+\tau) &= \sum_{s=0}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} \sum_{j=1}^{s+1} P_{s,j} j \tau^{j-1} \\ &= \sum_{s=0}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} \Big(\tau^s + \sum_{j=1}^{s} P_{s,j} j \tau^{j-1}\Big) \\ &= x(t_\lambda+\tfrac{\tau}{\lambda}) + R_\lambda(\tau), \end{aligned} \tag{3.38}$$

where

$$\begin{aligned} R_\lambda(\tau) &= \sum_{s=0}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} \sum_{j=1}^{s} P_{s,j} j \tau^{j-1} \\ &= \sum_{j=1}^{\infty} j\tau^{j-1} \sum_{s=j}^{\infty} \frac{x^{(s)}(t_\lambda)}{\lambda^s s!} P_{s,j} \\ &=: \sum_{j=1}^{\infty} j\tau^{j-1} b_j. \end{aligned} \tag{3.39}$$

A similar estimate for b_j is

$$|b_j| < \frac{2}{j!}\Big(\frac{\pi}{\lambda}\Big)^{j}, \tag{3.40}$$

which, through (3.38) and (3.39), provides us with the estimate

$$|X''_\lambda(N_\lambda+\tau) - \tfrac{1}{\lambda}x'(t_\lambda+\tfrac{\tau}{\lambda})| \leq 2(\tfrac{\pi}{\lambda})^2 e^{\tau\pi/\lambda}. \tag{3.41}$$

Now,

$$|x'(t_\lambda+\tfrac{\tau}{\lambda}) - x'(t)| \leq \frac{(\tau+1)}{\lambda}\pi^2, \tag{3.42}$$

so that

$$\frac{1}{2}|x'(t)| \le |x'(t_\lambda + \tfrac{\tau}{\lambda})| \le \frac{3}{2}|x'(t)|, \tag{3.43}$$

for all $0 \le \tau \le C\lambda|x'(t)|$, where C is a sufficiently small absolute constant. Hence, from (3.41) and (3.43), it follows that

$$\begin{aligned} |X_\lambda''(N_\lambda + \tau)| &\ge \frac{1}{\lambda}|x'(t_\lambda + \tfrac{\tau}{\lambda})| - 2e^{\pi\tau/\lambda}(\tfrac{\pi}{\lambda})^2 \\ &\ge C_1\frac{|x'(t)|}{\lambda} \end{aligned} \tag{3.44}$$

if $\lambda \ge C'|x'(t)|^{-1}$. It follows similarly that

$$|X_\lambda''(N_\lambda + \tau)| \le C_2\frac{|x'(t)|}{\lambda} \tag{3.45}$$

for the same range of λ and τ. □

Let us now apply Theorem 3.6 with $f = kX_\lambda$, $a = N_\lambda + 1$, and $b = N_\lambda + n$, and assume $n \le C_3|x'(t)|\lambda$. It follows from (3.31) that the number of integral terms in the right hand side of (3.13) is bounded by

$$\begin{aligned} |\beta - \alpha + 3| &\le 3 + k(n-1)\sup_{1\le\tau\le n}|X_\lambda''(N_\lambda + \tau)| \\ &\le 3 + k(n-1)C_2\frac{|x'(t)|}{\lambda}. \end{aligned} \tag{3.46}$$

On the other hand, using Theorem 3.7 (for $r = 2$) together with (3.31), each exponential integral term in (3.13) is bounded by $C(k|x'(t)|/\lambda)^{-1/2}$. Combining this with the bound on the number of terms that we have just found, we get

$$|S_{\lambda,k}(n)| \le C_1 n\Big(\frac{k}{\lambda}\Big)^{1/2} + C_2\Big(\frac{k}{\lambda}\Big)^{-1/2}|x'(t)|^{-1/2} + O(\log(2+k)), \tag{3.47}$$

where in the first term we have made use of $\|x'\|_{L^\infty} \le \pi$ (which follows from Bernstein's inequality), and in the logarithmic term, of $|\beta - \alpha| \le k$ for the given range of n. Note that, for small k, this bound significantly improves the trivial bound n. Now, if in (3.29), one chooses $K \sim \lambda^{1/3}$, then (3.47) yields our desired estimate (3.28).

4. We finish the proof of Theorem 3.23 by bounding (3.26) for a particular family of filters which we construct next. For this, we fix a filter φ such that $\hat\varphi$ is C^∞, $\mathrm{supp}(\hat\varphi) \subset [-c_0\pi, c_0\pi]$ for some small fixed $c_0 > 1$, and $\hat\varphi(\xi) = 1$ on $[-\pi, \pi]$. Then φ is a Schwartz function, i.e., φ has rapidly decreasing derivatives: there are constants $C_N^{(l)}$ for all $N \ge 0$ and $l \ge 0$ such that

$$|\varphi^{(l)}(t)| \le \frac{C_N^{(l)}}{(1+|t|)^N}. \tag{3.48}$$

For a small $\eta > 0$, we set $\Omega_\lambda = \lambda^{\eta/2}$ and define φ_λ by

$$\varphi_\lambda(t) = \Omega_\lambda\varphi(\Omega_\lambda t) \tag{3.49}$$

for $\lambda \geq 1$. Then $\hat{\varphi}_\lambda(\xi) = \hat{\varphi}(\xi/\Omega_\lambda)$ and hence $\{\varphi_\lambda\}$ is an admissible family of reconstruction filters. We return to the expression (3.26). For small n (i.e., for $|n| \leq c|x'(t)|\lambda$ for a sufficiently small constant c), we will use the estimate (3.28) in the form $O(|x'(t)|^{-1/2}\lambda^{2/3})$, and for large n, the trivial estimate $|n|/2$. Thus,

$$\begin{aligned} |e_\lambda(t)| &\leq O(|x'(t)|^{-1/2}\lambda^{2/3})\frac{1}{\lambda}\sum_{|n|\leq c|x'(t)|\lambda} |\Delta^2_{1/\lambda}\varphi_\lambda(-\tfrac{n}{\lambda}+\delta_\lambda)| \\ &\quad + \frac{1}{\lambda}\sum_{|n|>|x'(t)|c\lambda} \frac{|n|}{2}|\Delta^2_{1/\lambda}\varphi_\lambda(-\tfrac{n}{\lambda}+\delta_\lambda)|. \end{aligned} \tag{3.50}$$

The sum in the first term can easily be bounded by

$$2\lambda^{-1}\|\varphi''_\lambda\|_{L^1} = 2\lambda^{-1+\eta}\|\varphi''\|_{L^1}, \tag{3.51}$$

and the sum in the second term by

$$\sum_{|n|>c|x'(t)|\lambda} \frac{|n|}{2}\cdot\frac{2}{\lambda^2}\cdot\frac{\Omega_\lambda^3 C_N^{(2)}}{(1+\Omega_\lambda|n|\lambda^{-1})^N} \leq C(N,|x'(t)|)\Omega_\lambda^{-N+3}, \tag{3.52}$$

for all N. We choose N such that $(N-3)\eta/2 > 1/3$. Combining (3.50), (3.51), and (3.52) results in the estimate

$$|e_\lambda(t)| \leq C'(\eta,|x'(t)|)\lambda^{-4/3+\eta}, \tag{3.53}$$

concluding the proof. □

Remarks and an experiment: The key estimate in obtaining the error bound in this theorem was the estimate (3.47) for the exponential sums $S_{\lambda,k}(n)$. Therefore better estimates for these exponential sums can potentially yield improvements of the exponent of λ we have obtained to be $-4/3+\eta$. To test this, we can examine the growth of these exponential sums numerically. Noting that the dependence on n for these estimates did not really matter much in our analysis, we smooth out the role of n by looking at the behavior of

$$\sup_{1\leq n\leq c|x'(t)|\lambda} |S_{\lambda,k}(n)|$$

as we fix λ and vary k. However, this quantity still has a wild behavior in k. Reconsidering the role of k in the Erdős-Turán inequality (3.29), we look at the maximal value

$$\mathcal{M}_\lambda(K_0) := \sup_{1\leq k\leq K_0}\ \sup_{1\leq n\leq c|x'(t)|\lambda} |S_{\lambda,k}(n)| \tag{3.54}$$

for a fixed K_0, whose value is to be set depending on λ. We must then consider the quantity $\lambda/K_0 + \mathcal{M}_\lambda(K_0)\log(K_0)$ in (3.29). Clearly, the behavior of this quantity is determined by the asymptotic behavior of $\mathcal{M}_\lambda(K_0)$ for large K_0; in order to lead to any improvement in the estimate (3.28), the "optimal" K_0 should be larger than

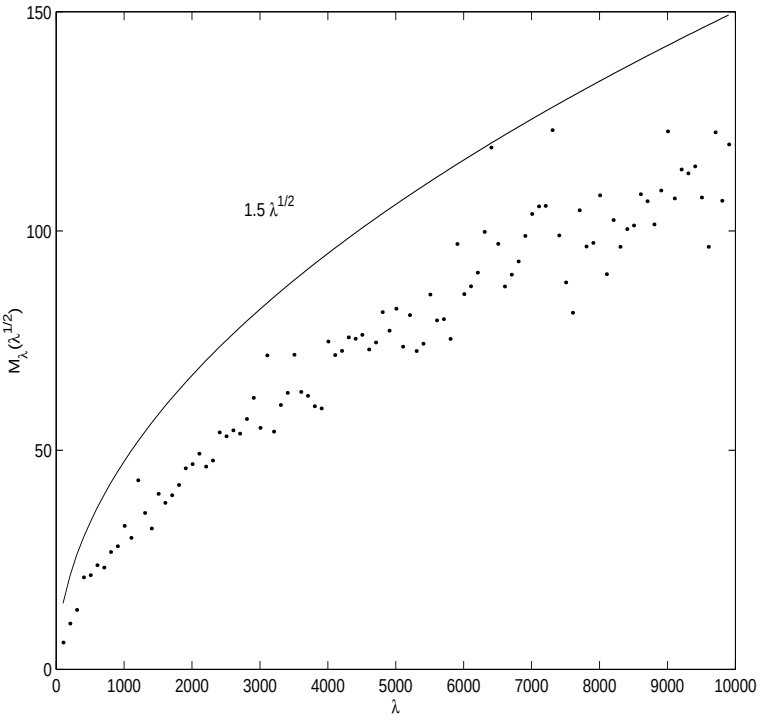

Figure 3.1: Plot of $\mathcal{M}_\lambda(\lambda^{1/2})$ vs. λ for $x(t) = \sin(t)$ and around $t = 1$. The solid curve is $1.5\,\lambda^{1/2}$. The exponential sums $S_{\lambda,k}(n)$ were computed for $n \leq 0.3\lambda$.

$\lambda^{1/3}$. On the other hand, certainly we would not want to set it too large. Since the empirical estimates for $|e_\lambda|$ (which are usually computed in various average senses) are of the order $\lambda^{-3/2}$, let's set $K_0 \sim \lambda^{1/2}$. For K_0 this large, the bound (3.47) certainly gets worse, giving only the naive $O(\lambda^{3/4})$. However, if the true behavior of the exponential sums is better than indicated by (3.47), we may do better. A numerical behavior of $\mathcal{M}_\lambda(\lambda^{1/2})$ of smaller order than $O(\lambda^{3/4})$ would be an indicator of possible improvement. However, we would still need to beat $O(\lambda^{2/3})$ for an improvement.

To test this approach, we employ the function $x(t) = \sin(t)$ and consider the point $t = 1$, at which the derivative $x'(1) = \cos(1) \approx 0.5403$ is nonvanishing and not too small. For λ in the range $[1, 10000]$, we compute $\mathcal{M}_\lambda(\lambda^{1/2})$. The result is an $O(\lambda^{1/2})$ behaviour, as plotted in Figure 3.1. If true, this would imply that $|e_\lambda(1)| = O(\lambda^{-3/2})$ for this function. This is a good sign that with a more refined analysis of the exponential sums $S_{\lambda,k}(n)$, it may be possible to achieve a similar bound for arbitrary bandlimited functions. In fact, further numerical experiments indicate that $\mathcal{M}_\lambda(K_0)$ does not depend significantly on K_0. One can conjecture the bound

$$\mathcal{M}_\lambda(K_0) = O(\lambda^{1/2} \log^\gamma K_0) \tag{3.55}$$

for some $\gamma > 0$.

Let us point out a possible strategy to improve our bound towards this conjecture. In Figure 3.2, we have plotted the sums $\{S_{\lambda,k}(n) : n = 1, \ldots, c|x'(t)|\lambda\}$ in the complex

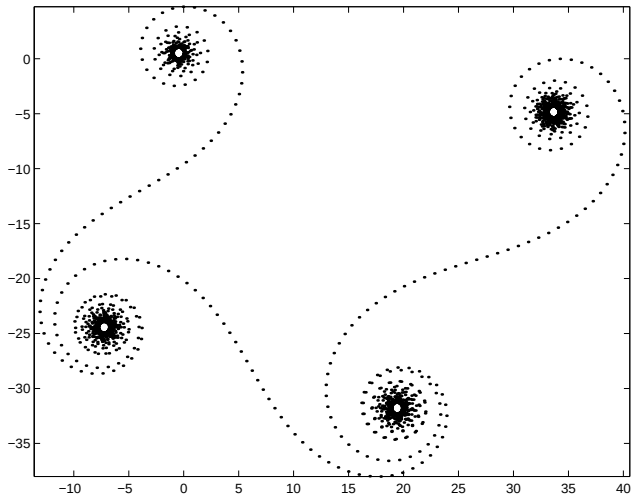

Figure 3.2: Plot of $\{S_{\lambda,k}(n) : n = 1, \ldots, 0.3\lambda\}$ for $x(t) = \sin(t)$, around $t = 1$. $\lambda = 10000$, $k = 30$.

plane for the same sinusoidal function given above, and for a fixed λ and k. We have set $t = 1$, $\lambda = 10000$, $c|x'(t)| = 0.3$, and $k = 30$. Remember that our method for estimating these sums had two ingredients: The truncated Poisson formula (3.13), and van der Corput estimate (3.14). The first one basically decomposes the graph into its "arms" where each arm starts from the center of a spiral and ends at the next center. (There are 3 arms in Figure 3.2.) Then the length of each arm is estimated using (3.14). However, this method does not really take into account the phase cancellations due to the relative orientations of these arms. This can be important when the number of arms is large. In Figure 3.3, we have $k = 350$, leading to many arms; their orientation changes sufficiently from one part to the other to contain the whole figure in a small square.

The estimates using the truncated Poisson summation formula can be refined by computing the phase of each integral. The resulting method is called the "Process B" of van der Corput [15, p. 54]. This method amounts to solving the roots of the equation $f'(x) = \nu$, for integers $\nu \in [f'(a), f'(b)]$, where f is the phase function for the exponential sum $S = \sum_a^b e^{2\pi i f(n)}$. This method is certainly powerful for functions that have explicit analytic forms. But given the generality assumed for the functions we consider, it becomes much more challenging to apply this method to our case. However, we believe that further analysis in this direction will improve the bounds.

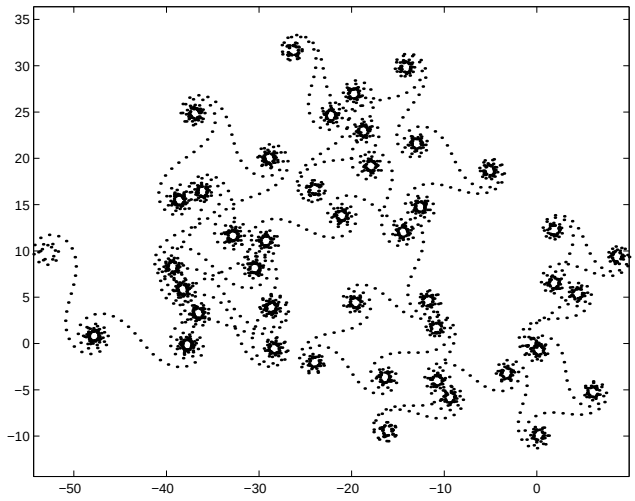

Figure 3.3: The same plot as in Figure 3.2, but for $k = 350$.

3.3 Second Order Systems

A second order scheme satisfies the difference equation

$$\Delta^2 u(n) = x(n) - q(n), \tag{3.56}$$

where x is a sequence in $[0, 1]$ and q is the output bit sequence in $\{0, 1\}^{\mathbb{Z}}$. The value of each $q(n)$ is computed by applying some nonlinear rule $Q(\cdot)$ to finite collections of previous values $\{u(n-1), u(n-2), \ldots\}$ of the state variable u, and previous values $\{x(n), x(n-1), \ldots\}$ of the input x, so that the difference equation (3.56) can be solved by recursion. In our discussion, we shall consider schemes in which $q(n)$ depends on $u(n-1), u(n-2)$, and $x(n)$ only. By defining

$$v(n) = \Delta u(n), \tag{3.57}$$

(3.56) can be rewritten in the following canonical form:

$$\begin{pmatrix} v(n) \\ u(n) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v(n-1) \\ u(n-1) \end{pmatrix} + (x(n) - q(n)) \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{3.58}$$

We will use the short-hand notation

$$\mathbf{u} = \begin{pmatrix} v \\ u \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{e} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{3.59}$$

so that (3.58) reads

$$\mathbf{u}(n) = \mathbf{A}\,\mathbf{u}(n-1) + (x(n) - q(n))\mathbf{e}. \tag{3.60}$$

We set $q(n) = Q(\mathbf{u}(n), x(n))$ and define a partition $\{\Omega_x^0, \Omega_x^1\}$ of the plane by setting $\Omega_x^0 = \{\mathbf{u} : Q(\mathbf{u}, x) = 0\}$, and $\Omega_x^1 = \{\mathbf{u} : Q(\mathbf{u}, x) = 1\}$. Hence, defining the

piecewise affine transformation $\mathbf{T}_x : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$\mathbf{T}_x\mathbf{u} = \begin{cases} \mathbf{A}\,\mathbf{u} + x\mathbf{e}, & \text{if } \mathbf{u} \in \Omega_x^0, \\ \mathbf{A}\,\mathbf{u} + (x-1)\mathbf{e}, & \text{if } \mathbf{u} \in \Omega_x^1, \end{cases} \tag{3.61}$$

the recursion (3.60) can be rewritten once again as

$$\mathbf{u}(n) = \mathbf{T}_{x(n)}\mathbf{u}(n-1). \tag{3.62}$$

In this setup, the whole variability of the second order schemes is reduced to the selection of the sets Ω_x^0 and Ω_x^1. In schemes used in engineering practice, these sets have been chosen to be half-spaces in $\mathbb{R}^2$, for then the computation of $q(n)$ in an electronic circuit is relatively easy since $q(n) = H(a_1u(n-1)+a_2u(n-2)+a_3x(n)+a_4)$, where $H = \chi_{[0,\infty)}$ is the Heaviside step function. We shall be interested in such half-space partitions as well as in schemes with more general partitions of the plane.

There are two types of questions regarding the transformation $\mathbf{T}_x$ and the associated dynamical system. The first is the question of stability; once stability is established, one can then immediately derive the basic decay estimate in λ, as we saw in Chapter 2. The second type of questions have to do with further properties of these dynamical systems, such as investigation of finer analytic and algebraic properties of the invariant sets of $\mathbf{T}_x$. As we shall see, improved estimates will depend on results in this direction.

3.3.1 Stable Schemes and Tiling Invariant Sets

We shall give two examples of stable second order sigma-delta schemes corresponding to two different types of "quantization rules". The first one will be a *linear* rule. For us, a general linear rule is given by

$$Q(\mathbf{u}, x) = \begin{cases} 0, & \text{if } \gamma v + u + \kappa(x) < 0, \\ 1, & \text{if } \gamma v + u + \kappa(x) \geq 0, \end{cases} \tag{3.63}$$

where γ is a real number and $\kappa(\cdot)$ is a real-valued function that controls the positioning of the line separating the corresponding half-spaces Ω_x^0 and Ω_x^1. Denote by V_y the vertical translation defined by $V_y(v,u) = (v, u+y)$. Let $\mathring{\mathbf{T}}_x$ be the transformation for which $\kappa \equiv 0$. It is easy to check that

$$\mathbf{T}_x = V_{\kappa(x)}^{-1}\mathring{\mathbf{T}}_x V_{\kappa(x)}. \tag{3.64}$$

Hence, for constant inputs (i.e., $x(m) = x$ for all m), there is no loss of generality if we assume $\kappa \equiv 0$, since then one has

$$\mathbf{T}_x^{(n)} = V_{\kappa(x)}^{-1}\mathring{\mathbf{T}}_x^{(n)} V_{\kappa(x)}. \tag{3.65}$$

On the other hand, for variable input, there may be significant gains by employing a rule with nonzero κ. For instance, it is conjectured [19] that if κ is adjusted to fix the centroids of the invariant sets of $\mathbf{T}_x$ at the same point for all x, then this leads to a better decay of the approximation error.

#1. A particular linear rule ($\gamma = 2$)

In this example, we shall analyze the following particular rule in which $\gamma = 2$, and $\kappa \equiv 0$. That is,

$$\mathbf{T}_x(v,u) = \begin{cases} (v+x, v+u+x), & \text{if} \quad 2v+u < 0, \\ (v+x-1, v+u+x-1), & \text{if} \quad 2v+u \geq 0. \end{cases} \tag{3.66}$$

By the global stability of $\mathbf{T}_x$, we mean that all initial points in $\mathbb{R}^2$ eventually get trapped by a fixed bounded set (independent of x), in which they stay forever. This is actually stronger than what is sufficient to ensure that $(u(n))$ is bounded for an arbitrary sequence $(x(n))$ in $[0,1]$. For the latter, it is enough to have some bounded set Γ that is positively invariant under all $\mathbf{T}_x$, i.e. $\mathbf{T}_x(\Gamma) \subset \Gamma$, for all x. For constant inputs, the uniformity condition over x can be relaxed and these questions can be asked for individual values of x only.

We shall not analyze in detail the issue of global stability, but only note that the following Lyapunov function can be used to prove this property for the linear rule we have described (and actually, for all $\gamma > 1/2$):

$$\mathcal{L}_x(v,u) = \begin{cases} x(-2u+v) + v^2, & \text{if} \quad 2v+u < 0, \\ (1-x)(2u-v) + v^2, & \text{if} \quad 2v+u \geq 0. \end{cases} \tag{3.67}$$

In the following discussion, we give an explicit parametrization of polygons $\mathcal{P}_x$ that stay invariant under the transformation $\mathbf{T}_x$, for x in the range $[\frac{1}{3}, \frac{2}{3}]$. Note that for all $\mathbf{u} \notin \partial\Omega_x^0$, one has $\mathbf{T}_{1-x}(\mathbf{u}) = -\mathbf{T}_x(-\mathbf{u})$, so that it suffices[3] to consider only the range $[\frac{1}{2}, \frac{2}{3}]$.

For each $x \in (\frac{1}{2}, \frac{2}{3}]$, we define

$$k_x = \left\lceil \frac{1}{2}\sqrt{1 + \frac{1}{2x-1}} \right\rceil, \tag{3.68}$$

and for $k = 1, 2, \ldots$,

$$\alpha_k = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{4k^2 - 1}. \tag{3.69}$$

One easily checks that $k_x = 1$ for $x = \alpha_1 = 2/3$, and $k_x = k$ for $x \in [\alpha_k, \alpha_{k-1})$, where $k \geq 2$. We also note that $(\alpha_k)_{k=1}^{\infty}$ is a monotonically decreasing sequence that converges to $1/2$.

Given $x \in [\alpha_k, \alpha_{k-1})$, we define the polygon $\mathcal{P}_x$ with the following set of vertices listed in the counter-clockwise direction (see Figure 3.4):

$$\{O_1, P_k, \ldots, P_1, Q_k, \ldots, Q_1, R_3, R_2, R_1, S_1, \ldots, S_k, S_a, T_1, \ldots, T_{k+1}\}, \tag{3.70}$$

where these points are parametrically defined as in the following list:

[3]The small exception for the points $\mathbf{u}$ on the line $2v + u = 0$ does not affect the discussion.

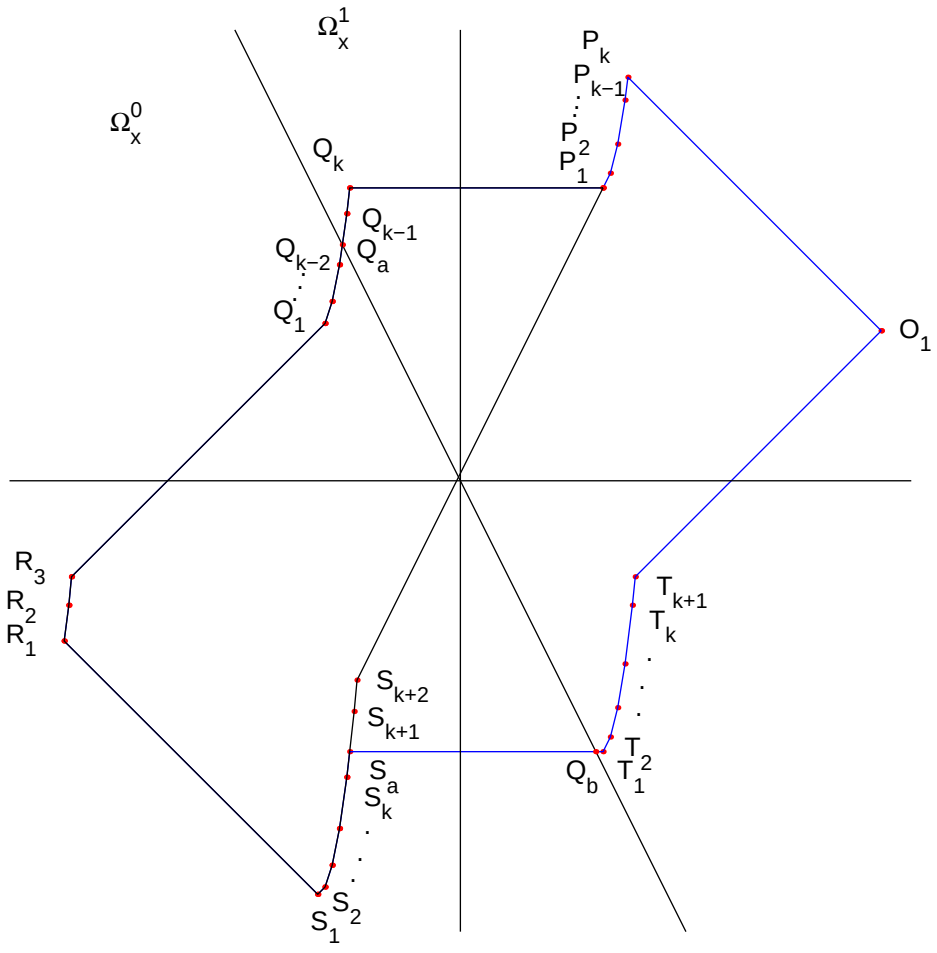

Figure 3.4: The invariant set for the dynamical system given in (3.66) for an arbitrary k value. (In this figure, $k = 5$ and $x \in [\alpha_5, \alpha_4)$.)

$$
\begin{aligned}
O_1 &= \begin{pmatrix} 1-0.5x \\ -1+2.5x \end{pmatrix} \\
P_k &= \begin{pmatrix} \frac{1}{4}+\frac{1}{4(2k-1)}+(x-0.5)(k-2) \\ \frac{3}{4}-\frac{1}{4(2k-1)}-(x-0.5)(k-4) \end{pmatrix} \\
P_j &= \begin{pmatrix} \frac{1}{4}+(x-0.5)(2j-1.5) \\ \frac{1}{2}+(x-0.5)(2j^2-2j+3) \end{pmatrix}, \qquad j=1,\ldots,k-1 \\
Q_k &= \begin{pmatrix} -\frac{1}{4}+\frac{1}{4(2k-1)}+(x-0.5)(k-1) \\ 3x-1 \end{pmatrix} \\
Q_j &= \begin{pmatrix} -\frac{1}{4}+(x-0.5)(2j-0.5) \\ \frac{1}{4}+(x-0.5)(2j^2+2.5) \end{pmatrix}, \qquad j=1,\ldots,k-1 \\
Q_a &= \begin{pmatrix} -\frac{1}{4}+\frac{1}{4(2k-1)}+(x-0.5)(k-3) \\ \frac{1}{2}-\frac{1}{2(2k-1)}-(2x-1)(k-3) \end{pmatrix}, \\
R_3 &= \begin{pmatrix} -\frac{3}{4}+\frac{1}{4(2k-1)}+(x-0.5)k \\ -\frac{1}{4}+\frac{1}{4(2k-1)}+(x-0.5)(k+3) \end{pmatrix} \\
R_2 &= \begin{pmatrix} -\frac{3}{4}+(x-0.5)(2k-1.5) \\ -\frac{1}{2}+(x-0.5)(2k^2-2k+3) \end{pmatrix} \\
R_1 &= \begin{pmatrix} -\frac{3}{4}+\frac{1}{4(2k-1)}+(x-0.5)(k-2) \\ -\frac{1}{4}-\frac{1}{4(2k-1)}-(x-0.5)(k-4) \end{pmatrix} \\
S_{j+1} &= \begin{pmatrix} -\frac{1}{4}+(x-0.5)(2j-0.5) \\ -\frac{3}{4}+(x-0.5)(2j^2+2.5) \end{pmatrix}, \qquad j=0,\ldots,k-1 \\
S_a &= \begin{pmatrix} -\frac{1}{4}+\frac{1}{4(2k-1)}+(x-0.5)(k-1) \\ 3x-2 \end{pmatrix} \\
Q_b &= \begin{pmatrix} -1.5x+1 \\ 3x-2 \end{pmatrix} \\
T_j &= \begin{pmatrix} \frac{1}{4}+(x-0.5)(2j-1.5) \\ -\frac{1}{2}+(x-0.5)(2j^2-2j+3) \end{pmatrix}, \qquad j=1,\ldots,k \\
T_{k+1} &= \begin{pmatrix} \frac{1}{4}+\frac{1}{4(2k-1)}+(x-0.5)k \\ -\frac{1}{4}+\frac{1}{4(2k-1)}+(x-0.5)(k+3) \end{pmatrix}.
\end{aligned}
$$

It is straightforward (albeit lengthy) to verify that, under the action of $\mathbf{T}_x$, these vertices are mapped onto each other as follows: The first group is given by

$$
\begin{aligned}
Q_b &\longmapsto S_1 \\
T_j &\longmapsto S_{j+1}, \quad j=1,\ldots,k+1, \\
O_1 &\longmapsto P_1 \\
P_j &\longmapsto Q_j, \quad j=1,\ldots,k, \\
Q_k &\longmapsto R_3
\end{aligned}
$$

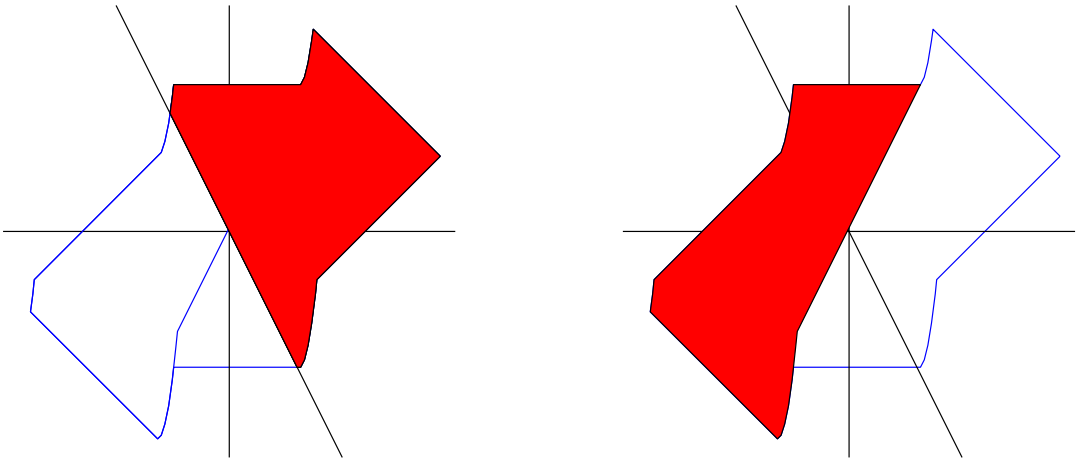

Figure 3.5: A schematic diagram of the action of $\mathbf{T}_x$ on Γ_x.

$$\begin{aligned} Q_{k-1} &\longmapsto R_2 \\ Q_a &\longmapsto R_1, \end{aligned}$$

which corresponds, in Figure 3.5, to the mapping of the shaded polygon on the left to the shaded polygon on the right.
The second group is given similarly by

$$\begin{aligned} Q_a &\longmapsto P_k \\ Q_j &\longmapsto P_{j+1}, \quad j = 1, \ldots, k-2, \\ R_3 &\longmapsto S_{k+2} \\ R_2 &\longmapsto S_{k+1} \\ R_1 &\longmapsto S_a \\ S_j &\longmapsto T_j, \quad j = 1, \ldots, k, \\ S_a &\longmapsto T_{k+1} \\ Q_b &\longmapsto O_1, \end{aligned}$$

where for Q_a and Q_b, the mapping is understood in the sense that

$$\lim_{\substack{P \to Q_a \\ P \in \Omega_x^0}} \mathbf{T}_x(P) = P_k, \qquad \text{and} \qquad \lim_{\substack{P \to Q_b \\ P \in \Omega_x^0}} \mathbf{T}_x(P) = O_1. \tag{3.71}$$

Hence, it follows, due to the affinity of $\mathbf{T}_x$ on each half-space Ω_x^0 and Ω_x^1, that the set Γ_x, defined to be the interior of the polygon $\mathcal{P}_x$, is mapped onto itself under the action of $\mathbf{T}_x$.

#2. The quadratic rule of N. Thao

We saw above that although a linear quantization rule is simple, it leads to complicated invariant sets which are somewhat cumbersome to analyze. On the other hand,

the following scheme designed by N. Thao leads to much simpler invariant sets. From a topological view point, these invariant sets are equivalent to a square: each of them is a simply-connected set with a piecewise quadratic boundary that is composed of four pieces. This follows from a clever choice of the partition $\{\Omega_x^0, \Omega_x^1\}$ where the boundaries of these sets are particular parabolas. To construct the scheme, we shall first make a change of coordinates:

$$\tilde{\mathbf{u}} = \Phi_x(\mathbf{u}), \tag{3.72}$$

where Φ_x is a bijection to be specified soon. Let $\tilde{\mathbf{T}}_x$ denote the transformation in the new coordinate system, i.e.,

$$\tilde{\mathbf{T}}_x \Phi_x = \Phi_x \mathbf{T}_x. \tag{3.73}$$

We shall seek to define the transformation Φ_x so that

$$\Phi_x(\mathbf{T}_x \mathbf{u}) = \Phi_x(\mathbf{u}) + (x-1)\mathbf{f}, \quad \text{for all } \mathbf{u} \in \Omega_x^1, \tag{3.74}$$

for some fixed $\mathbf{f}$, ensuring that the transformation $\tilde{\mathbf{T}}_x$ would reduce to a translation by $(x-1)\mathbf{f}$ on $\tilde{\Omega}_x^1 := \Phi_x(\Omega_x^1)$. This can be achieved by a shift of the ordinate u by an amount quadratic in v, given by

$$\begin{aligned} (\tilde{v}, \tilde{u}) &:= \Phi_x(v, u) \\ &:= (v, u - \alpha(v+\beta)^2), \end{aligned} \tag{3.75}$$

where α and β are appropriately chosen. The simple choice of a horizontal translation (for which $\mathbf{f} = (1 \;\; 0)^T$) yields the values $\alpha = \frac{1}{2(x-1)}$ and $\beta = \frac{x-1}{2}$.

The transformation $\tilde{\mathbf{T}}_x$ is still affine on $\tilde{\Omega}_x^0 := \Phi_x(\Omega_x^0)$. A straightforward computation gives

$$\tilde{\mathbf{T}}_x \tilde{\mathbf{u}} = \mathbf{A}_x \tilde{\mathbf{u}} + x\mathbf{g}_x, \quad \text{for all } \tilde{\mathbf{u}} \in \tilde{\Omega}_x^0, \tag{3.76}$$

where

$$\mathbf{A}_x = \begin{pmatrix} 1 & 0 \\ -\frac{1}{x-1} & 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{g}_x = \begin{pmatrix} 1 \\ -\frac{1}{2(x-1)} \end{pmatrix}. \tag{3.77}$$

The final ingredient is the specification of the partition $\{\Omega_x^0, \Omega_x^1\}$, or equivalently the partition $\{\tilde{\Omega}_x^0, \tilde{\Omega}_x^1\}$. This is given, for $1/2 \le x < 1$, by

$$\tilde{\Omega}_x^1 = \{(\tilde{v}, \tilde{u}) : \tilde{u} \ge \frac{\tilde{v}}{x-1} + C_x\} \tag{3.78}$$

for a suitable constant C_x. (This constant can be arbitrary, but we shall choose its value to be $(1.5x-1)/(x-1)$ for normalization purposes.) To summarize, $\tilde{\mathbf{T}}_x$ is given by

$$\tilde{\mathbf{T}}_x(\tilde{v}, \tilde{u}) = \begin{cases} (\tilde{v} + x, -\tilde{v}/(x-1) + \tilde{u} - x/2(x-1)), & \text{if } \; \tilde{u} - \tilde{v}/(x-1) < C_x, \\ (\tilde{v} + x - 1, \tilde{u}), & \text{if } \; \tilde{u} - \tilde{v}/(x-1) \ge C_x. \end{cases} \tag{3.79}$$

For $1/2 \le x < 1$ and the given choice of C_x, the invariant set $\tilde{\Gamma}_x$ of $\tilde{\mathbf{T}}_x$ turns out to be a trapezoid whose vertices $\{\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4\}$ are given by $\tilde{P}_1 = (1 - 0.5x, 1)$,

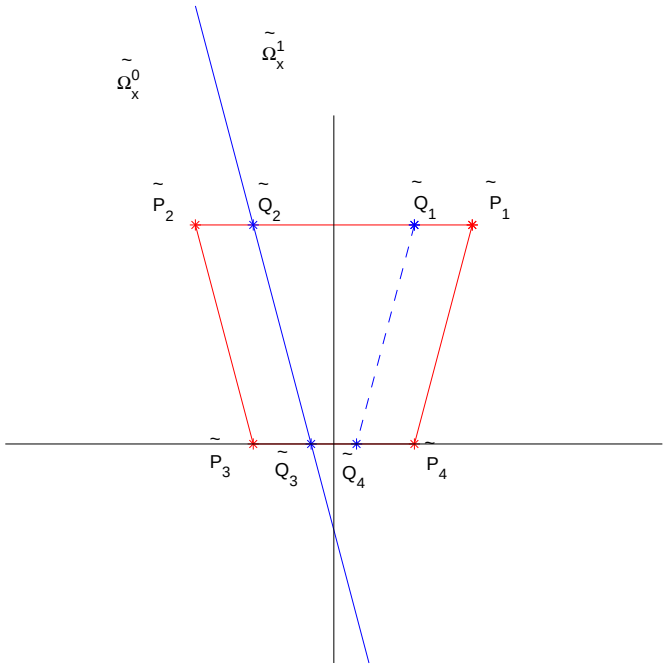

Figure 3.6: The invariant set in the transformed domain. ($x \sim 0.7357$)

$\tilde{P}_2 = (-1+0.5x, 1)$, $\tilde{P}_3 = (-0.5x, 0)$, $\tilde{P}_4 = (0.5x, 0)$. Let us also define $\tilde{Q}_1 = (0.5x, 1)$, $\tilde{Q}_2 = (-0.5x, 1)$, $\tilde{Q}_3 = (-1.5x + 1, 0)$, $\tilde{Q}_4 = (1.5x - 1, 0)$. The invariance of $\tilde{\Gamma}_x$ under $\tilde{\mathbf{T}}_x$ is easily verified by checking that the trapezoid $\tilde{P}_1\tilde{Q}_2\tilde{Q}_3\tilde{P}_4$ is mapped to the trapezoid $\tilde{Q}_1\tilde{P}_2\tilde{P}_3\tilde{Q}_4$ and the parallelogram $\tilde{Q}_2\tilde{P}_2\tilde{P}_3\tilde{Q}_3$ is mapped to the parallelogram $\tilde{Q}_1\tilde{Q}_4\tilde{P}_4\tilde{P}_1$. (Note that while the first mapping is a pure translation in the horizontal direction, the second is a shear followed by a translation in both directions. See Figure 3.6.)

Let Γ_x be the corresponding invariant set of the original transformation $\mathbf{T}_x$. Then, it is clear that Γ_x will have a boundary that is composed of four parabolic pieces. This is illustrated in Figure 3.7.

For $0 \leq x < 1/2$, on the other hand, one simply defines $\mathbf{T}_x$ as $\mathbf{T}_x(\mathbf{u}) = -\mathbf{T}_{1-x}(-\mathbf{u})$ by setting $\Omega_x^0 := -\Omega_{1-x}^1$, and $\Omega_x^1 := -\Omega_{1-x}^0$.

3.3.2 Improved Estimates for Constant Inputs

The following proposition summarizes some of the properties of the dynamical systems given in §3.3.1. These properties will be used in this section to prove the error estimates regarding second order $\Sigma\Delta$ systems.

Proposition 3.13. *For each of the dynamical systems #1 and #2 given in §3.3.1, there exists a subinterval I of $[0, 1]$ such that for each $x \in I$, the map $\mathbf{T}_x$ possesses an invariant set Γ_x with the following properties:*

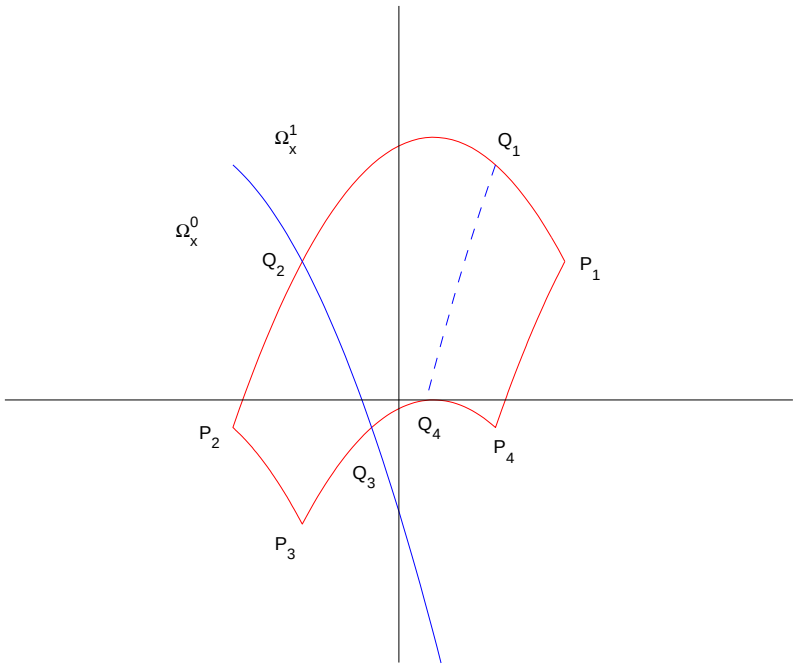

Figure 3.7: The invariant set in the original domain. ($x \sim 0.7357$)

1. $\mathbf{T}_x(\Gamma_x) = \Gamma_x$,

2. *There exists a positive constant* M_0 *such that*

$$\sup_{x \in I} \sup_{\mathbf{u} \in \Gamma_x} |\mathbf{u}| \leq M_0, \tag{3.80}$$

3. *There exists an absolute positive constant* C_0 *such that* $\Gamma_x \in \mathcal{M}_b$ *for all* $x \in I$, *where* $b(\epsilon) = C_0\epsilon$. *(see 3.1 for the definition of the family* $\mathcal{M}_b$.*)*

4. *For each* $x \in I$, *the set* Γ_x *is congruent to* $\mathbb{T}^2$ *modulo translations by vectors in* $\mathbb{Z}^2$, *up to possibly a set of measure zero. I.e., the translates of* Γ_x *by the integer lattice tile the plane:*

$$\Gamma_x + \mathbb{Z}^2 = \mathbb{R}^2, \tag{3.81}$$

where the equality is (possibly) up to a null subset of $\mathbb{R}^2$.

Proof. All of these properties can be checked in a straightforward manner using the explicit descriptions of the invariant sets given in §3.3.1. □

We continue with the definition of a particular filter that we shall employ to improve the basic estimate we gave in Chapter 2. While this filter is not the only possibility for this purpose, its simplicity will suit our purposes. Given the positive integer λ, let the discrete filter g_λ be defined by $g_\lambda(n) = \frac{1}{\lambda}\chi_{[0,1)}(\frac{n}{\lambda})$ and set

$$h_\lambda = g_\lambda * g_\lambda * g_\lambda, \tag{3.82}$$

and $e_\lambda := e_{\lambda,x} := x - q * h_\lambda$, as before. As in (3.21), we define

$$\mathrm{MSE}'(e_\lambda, I) = \left\| \int_I |e_{\lambda,x}(\cdot)|^2 \, dx \right\|_{\ell^\infty} \tag{3.83}$$

for a subinterval I of $[0,1]$.

For each of the second order sigma-delta schemes given by the dynamical systems #1 and #2, we assume that for each $x \in I$, the initial condition $\mathbf{u}(0)$ is chosen from Γ_x and the sequence $u(n)$ is defined for $n < 0$ by running the recursion backwards. (Note that this is possible since the transformation $\mathbf{T}_x$ is a bijection on the invariant set Γ_x.) We also assume that the output bit sequences are filtered by the family $\{h_\lambda\}$ given in (3.82). The following theorem is our improved estimate for second order schemes:

Theorem 3.14. *Under the assumptions listed above, the mean square error defined by (3.83) satisfies the estimate*

$$MSE\,'(e_\lambda, I) \le C\lambda^{-9/2} \log^2 \lambda. \tag{3.84}$$

Proof. We will divide the proof into several steps.

1. The filter h_λ is piecewise quadratic in n, and supported on $\{0, \ldots, 3\lambda - 1\}$. Furthermore, $\sum g_\lambda(n) = 1$ implies that $\sum h_\lambda(n) = 1$ as well. Since we assume that x is constant, this reduces the error e_λ to $(x - q) * h_\lambda$ as before. Substituting (3.56) into this expression yields

$$\begin{aligned} e_\lambda &= \Delta^2 u * h_\lambda \\ &= u * \Delta^2 h_\lambda \\ &= u * g_\lambda * (\Delta g_\lambda) * (\Delta g_\lambda) \\ &= u * g_\lambda * \frac{1}{\lambda^2}(\delta_0 - 2\delta_\lambda + \delta_{2\lambda}). \end{aligned} \tag{3.85}$$

In the last step we have made use of the fact that $\Delta g_\lambda = \frac{1}{\lambda}(\delta_0 - \delta_\lambda)$, where δ_a denotes the sequence $\delta_a(n) = 1$ if $n = a$, and $\delta_a(n) = 0$ if $n \ne a$. Thus, (3.85) reads

$$e_\lambda(n) = \frac{1}{\lambda^3} \left(\sum_{m=0}^{\lambda-1} u(n-m) - 2 \sum_{m=\lambda}^{2\lambda-1} u(n-m) + \sum_{m=2\lambda}^{3\lambda-1} u(n-m) \right). \tag{3.86}$$

Define $F : \mathbb{R}^2 \to \mathbb{R}$ by $F(\mathbf{w}) := F(w_1, w_2) := w_2$, and for integers $a < b$, let

$$E(a,b) := \left| \frac{1}{b-a} \sum_{a < n \le b} F(\mathbf{u}(n)) - \int_{\Gamma_x} F(\mathbf{w}) \, d\mathbf{w} \right|, \tag{3.87}$$

where Γ_x is the attracting invariant set of $\mathbf{T}_x$ defined in §3.3.1. Thus, (3.86) leads to the following inequality:

$$|e_\lambda(n)| \le \frac{1}{\lambda^2} \left(E(n-\lambda, n) + 2E(n-2\lambda, n-\lambda) + E(n-3\lambda, n-2\lambda) \right). \tag{3.88}$$

2. Our next objective is to find a good bound for $E(a,b)$. The definition given in (3.87) suggests using Koksma's inequality in two dimensions (or more precisely *the Koksma-Hlawka inequality*). However, the domain of integration is Γ_x which, at the moment, does not seem to fit the set-up for this inequality. We also do not have a closed form expression for the sequence $\{\mathbf{u}(n)\}$. We shall overcome these two difficulties by using the properties listed in Proposition 3.13 and the auxiliary sequence $\langle\mathbf{u}\rangle$ whose elements are the fractional parts of $(\mathbf{u}(n))_{n\in\mathbb{Z}}$.

For $\mathbf{v}=(v_1,v_2)\in\mathbb{Z}^2$, define $\Gamma_x(\mathbf{v})=\Gamma_x\cap[v_1,v_1+1)\times[v_2,v_2+1)$. This results in a partition of Γ_x as

$$\Gamma_x=\bigcup_{\mathbf{v}\in\mathcal{V}}\Gamma_x(\mathbf{v}) \tag{3.89}$$

for some finite subset $\mathcal{V}$ of $\mathbb{Z}^2$, and (including the possibility of having empty sets in the partition) uniformly in x, due to property 2 listed in Proposition 3.13. The tiling property of Γ_x results in the equality

$$\bigcup_{\mathbf{v}\in\mathcal{V}}[\Gamma_x(\mathbf{v})-\mathbf{v}]=[0,1)^2. \tag{3.90}$$

(One may also write $\langle\Gamma_x(\mathbf{v})\rangle$ for $\Gamma_x(\mathbf{v})-\mathbf{v}$.) Since F is linear, one has

$$\int_{\Gamma_x(\mathbf{v})}F(\mathbf{w})\,d\mathbf{w}=\int_{\Gamma_x(\mathbf{v})-\mathbf{v}}F(\mathbf{w})\,d\mathbf{w}+F(\mathbf{v})\int_{\Gamma_x(\mathbf{v})-\mathbf{v}}d\mathbf{w}, \tag{3.91}$$

which, after summing over all $\mathbf{v}\in\mathcal{V}$, results in

$$\int_{\Gamma_x}F(\mathbf{w})\,d\mathbf{w}=\int_{[0,1)^2}F(\mathbf{w})\,d\mathbf{w}+\sum_{\mathbf{v}\in\mathcal{V}}F(\mathbf{v})\int_{\Gamma_x(\mathbf{v})-\mathbf{v}}d\mathbf{w}. \tag{3.92}$$

Because of the tiling property, $\mathbf{u}(n)\in\Gamma_x(\mathbf{v})$ if and only if $\langle\mathbf{u}(n)\rangle\in\Gamma_x(\mathbf{v})-\mathbf{v}$. Hence

$$\begin{aligned}\langle\mathbf{u}(n)\rangle &= \mathbf{u}(n)-\sum_{\mathbf{v}\in\mathcal{V}}\mathbf{v}\;\chi_{\Gamma_x(\mathbf{v})}(\mathbf{u}(n))\\ &= \mathbf{u}(n)-\sum_{\mathbf{v}\in\mathcal{V}}\mathbf{v}\;\chi_{\Gamma_x(\mathbf{v})-\mathbf{v}}(\langle\mathbf{u}(n)\rangle),\end{aligned} \tag{3.93}$$

so that

$$\sum_{a<n\le b}F(\mathbf{u}(n))=\sum_{a<n\le b}F(\langle\mathbf{u}(n)\rangle)+\sum_{\mathbf{v}\in\mathcal{V}}F(\mathbf{v})\sum_{a<n\le b}\chi_{\Gamma_x(\mathbf{v})-\mathbf{v}}(\langle\mathbf{u}(n)\rangle) \tag{3.94}$$

Hence, it follows from (3.92) and (3.94) that

$$\begin{aligned}E(a,b) \;\le\; &\left|\frac{1}{b-a}\sum_{a<n\le b}F(\langle\mathbf{u}(n)\rangle)-\int_{[0,1)^2}F(\mathbf{w})\,d\mathbf{w}\right|\\ &+\sum_{\mathbf{v}\in\mathcal{V}}|F(\mathbf{v})|\left|\frac{1}{b-a}\sum_{a<n\le b}\chi_{\Gamma_x(\mathbf{v})-\mathbf{v}}(\langle\mathbf{u}(n)\rangle)-\int_{\Gamma_x(\mathbf{v})-\mathbf{v}}d\mathbf{w}\right|\end{aligned} \tag{3.95}$$

Denote by $D_{(a,b]}(\langle \mathbf{u} \rangle)$, the discrepancy of the sequence of points $\{\langle \mathbf{u}(n) \rangle\}_{a<n\leq b}$, and by $D_{(a,b]}(\langle \mathbf{u} \rangle, H)$, the discrepancy of the same set of points with respect to the set H. (Note that, through $\mathbf{u}$, these quantities implicitly depend also on x.) Using the Koksma-Hlawka inequality [14, Theorem 1.14], we have

$$\left| \frac{1}{b-a} \sum_{a<n\leq b} F(\langle \mathbf{u}(n) \rangle) - \int_{[0,1)^2} F(\mathbf{u})\, d\mathbf{u} \right| \leq C D_{(a,b]}(\langle \mathbf{u} \rangle) \tag{3.96}$$

for some absolute constant C. (Note that in general $C = C(F)$, but F is fixed in our whole discussion.) Now, from Proposition 3.13 and the definition of F, we have $|F(\mathbf{v})| \leq |\mathbf{v}| \leq M_0$, so that

$$E(a,b) \leq C\, D_{(a,b]}(\langle \mathbf{u} \rangle) + M_0 \sum_{\mathbf{v}\in\mathcal{V}} D_{(a,b]}(\langle \mathbf{u} \rangle, \Gamma_x(\mathbf{v}) - \mathbf{v}). \tag{3.97}$$

The cardinality $\#\mathcal{V}$ is bounded by M_0^2 so that Proposition 3.13 (Property 3) and Theorem 3.4 lead us to the estimate

$$\begin{aligned} E(a,b) &\leq C\, D_{(a,b]}(\langle \mathbf{u} \rangle) + C_2\, M_0^3\, D_{(a,b]}(\langle \mathbf{u} \rangle)^{1/2} \\ &\leq (C + C_2\, M_0^3) D_{(a,b]}(\langle \mathbf{u} \rangle)^{1/2}. \end{aligned} \tag{3.98}$$

Considering (3.88), any non-trivial bound for the discrepancy $D_{(a,b]}(\langle \mathbf{u} \rangle)$ will thus lead to an improved bound for the error e_λ.

3. Next, we derive a closed form analytical formula for the sequence $\langle \mathbf{u} \rangle$ and use this formula to estimate the discrepancy $D_{(a,b]}(\langle \mathbf{u} \rangle)$ via the Erdős-Turán-Koksma inequality. On $\mathbb{T}^2$, the recursion relation (3.60) turns into the bijective transformation

$$\langle \mathbf{T}_x \mathbf{u} \rangle = \langle \mathbf{A}\, \mathbf{u} + x\mathbf{e} \rangle, \tag{3.99}$$

which can be iterated forwards and backwards to write a solution for $\langle \mathbf{u}(n) \rangle$:

$$\langle \mathbf{u}(n) \rangle = \begin{cases} \langle \mathbf{A}^n\, \mathbf{u}(0) + x(\sum_{i=0}^{n-1} \mathbf{A}^i)\mathbf{e} \rangle & \text{if } n \geq 0, \\ \langle \mathbf{A}^n\, \mathbf{u}(0) - x(\sum_{i=n}^{-1} \mathbf{A}^i)\mathbf{e} \rangle & \text{if } n < 0. \end{cases} \tag{3.100}$$

On the other hand, after evaluating these two expressions, one finds a single analytical formula for all $n \in \mathbb{Z}$:

$$\langle \mathbf{u}(n) \rangle = \begin{pmatrix} \langle v(n) \rangle \\ \langle u(n) \rangle \end{pmatrix} = \begin{pmatrix} \langle v(0) + nx \rangle \\ \langle nv(0) + u(0) + \frac{1}{2} n(n+1)x \rangle \end{pmatrix}. \tag{3.101}$$

Define, for $\mathbf{k} = (k_1, k_2) \in \mathbb{Z}^2$,

$$S_{(a,b]}(\mathbf{k}, x) := \frac{1}{b-a} \sum_{a<n\leq b} e^{2\pi i \mathbf{k}\cdot\langle \mathbf{u}(n) \rangle}, \tag{3.102}$$

where the dependence on x becomes explicit if the formula (3.101) is inserted in this expression. Using the periodicity of the exponential function, one can rewrite $S_{(a,b]}$ as

$$S_{(a,b]}(\mathbf{k}, x) = \frac{1}{b-a} \sum_{a<n\leq b} c_n e^{2\pi i d_n x}, \tag{3.103}$$

where $c_n = e^{2\pi i[v(0)k_1 + (u(0)+nv(0))k_2]}$ and $d_n = nk_1 + \frac{1}{2}n(n+1)k_2$. This quantity is initially defined only for $x \in I$, where I is one of the intervals defined by Proposition 3.13, since we have evaluated $\mathbf{u}$ only for this interval. Using (3.103), we extend the definition of $S_{(a,b]}(\mathbf{k}, x)$ to all $x \in \mathbb{T}$.

Note that $|c_n| = 1$ and $d_n \in \mathbb{Z}$ for all n. Since d_n is a quadratic polynomial in n, it can attain any given value at most twice. Hence, if $S_{(a,b]}(\mathbf{k}, x)$ is rewritten as a trigonometric polynomial in x with distinct frequencies, the amplitude of each frequency will be bounded by $2/(b-a)$, since $\max_{n,m} |c_n + c_m| \leq 2$. Also, there will be at most $b-a$ distinct frequencies. Thus, using Parseval's theorem, one easily bounds $\|S_{(a,b]}(\mathbf{k}, \cdot)\|_{L^2(\mathbb{T})}$ by

$$\|S_{(a,b]}(\mathbf{k}, \cdot)\|_{L^2(\mathbb{T})} \leq \frac{2}{\sqrt{b-a}}. \tag{3.104}$$

Now, for any positive integer K, Theorem 3.5 yields the estimate

$$D_{(a,b]}(\langle \mathbf{u} \rangle) \leq C \left(\frac{1}{K} + \sum_{0<\|\mathbf{k}\|_\infty \leq K} \frac{1}{r(\mathbf{k})} \left| S_{(a,b]}(\mathbf{k}, x) \right| \right), \tag{3.105}$$

so that

$$\int_I D_{(a,b]}(\langle \mathbf{u} \rangle)\, dx \;\leq\; C \inf_{K\geq 1} \left(\frac{1}{K} + \sum_{0<\|\mathbf{k}\|_\infty \leq K} \frac{1}{r(\mathbf{k})} \left\| S_{(a,b]}(\mathbf{k}, x) \right\|_{L^1(\mathbb{T})} \right). \tag{3.106}$$

Using the inequality $\|\cdot\|_{L^1(\mathbb{T})} \leq \|\cdot\|_{L^2(\mathbb{T})}$, the bound (3.104), and

$$\begin{aligned} \sum_{0<\|\mathbf{k}\|_\infty \leq K} \frac{1}{r(\mathbf{k})} &= 4 \left(\sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \frac{1}{k_1 k_2} + \sum_{k_1=1}^{K} \frac{1}{k_1} \right) \\ &= O(\log^2 K), \end{aligned} \tag{3.107}$$

one concludes that

$$\begin{aligned} \int_I D_{(a,b]}(\langle \mathbf{u} \rangle)\, dx &\leq C' \inf_{K\geq 1} \left(\frac{1}{K} + (b-a)^{-1/2} \log^2 K \right) \\ &\leq C''(b-a)^{-1/2} \log^2(b-a). \end{aligned} \tag{3.108}$$

4. The final step is to combine the results of steps **1**, **2**, and **3**. We first combine the bounds (3.98) and (3.108) to obtain

$$\int_I E(n-\lambda, n)^2\, dx \leq C\lambda^{-1/2} \log^2 \lambda \tag{3.109}$$

uniformly in n. It next follows from (3.88) and the Schwarz's inequality that

$$|e_{\lambda,x}(n)|^2 \leq 6\lambda^{-4}\left(E(n-\lambda,n)^2 + E(n-2\lambda,n-\lambda)^2 + E(n-3\lambda,n-2\lambda)^2\right). \tag{3.110}$$

Finally, we combine (3.109) and (3.110) to get

$$\left\| \int_I |e_{\lambda,x}(\cdot)|^2 \, dx \right\|_{\ell^\infty} \leq C\lambda^{-9/2}\log^2\lambda, \tag{3.111}$$

proving our assertion (3.84). □

Chapter 4

Other Results and Considerations

4.1 More on First Order $\Sigma\Delta$ Quantization

4.1.1 Optimal MSE Estimates for Constants

Consider the first order $\Sigma\Delta$ scheme given by (2.7)-(2.9), or equivalently by (2.10) and (2.11). We would also like to consider an arbitrary initial condition $u(0) = u_0 \in [0, 1)$. Then, for constant input x, the output bits $q(n)$ will be such that

$$Q(n) = q(1) + \ldots + q(n) = \lfloor u_0 + nx \rfloor. \tag{4.1}$$

As x varies in $[0, 1]$, the collection of outcomes $\mathbf{q}_x(N) := (q(1), \ldots, q(N))$ define a scalar quantizer whose threshold points (other than the points 0 and 1) are given by a "modified" Farey series $\mathcal{S}_N(u_0)$, where

$$\mathcal{S}_N(u_0) := \{(j - u_0)/n \ : \ \ j = 1, \ldots, n; \ n = 1, 2, \ldots, N\}. \tag{4.2}$$

We leave out the proof of this statement since it uses the same argument as in the case $u_0 = 0$. (see Chapter 2, page 10.)

In this section, we study how to give lower and upper bounds for the mean squared quantization error if u_0 is kept fixed while the constant input x is drawn uniformly from $[0, 1]$. For lower bounds we assume optimal decoding, where u_0 is known to the decoder. The optimal MSE quantizer is described using the map $Q_{opt}(x; u_0)$, which maps x to the midpoint of the interval J containing x, the endpoints of J being successive elements of the threshold points in $\mathcal{S}_N(u_0)$. The map $Q_{opt}(x; u_0)$ is the optimal quantizer under our assumption that the quantity x being quantized is uniformly distributed in $[0, 1]$ and is independent of u_0. Our objective is to lower bound the mean squared-error, given by the integral

$$MSE_{u_0}(Q_{opt}) := \int_0^1 (x - Q_{opt}(x; u_0))^2 dx, \tag{4.3}$$

uniformly in u_0. In the upper bound case, we suppose u_0 is fixed but unknown to the decoder, and we consider a specific decoding algorithm that uses the triangular

filter of length N. Let $Q_h(x;u_0)$ denote the estimate of x, using the filter $h = (h(1),\ldots,h(N))$. It is given by

$$Q_h(x;u_0) = \sum_{n=1}^{N} h(n)q(n). \tag{4.4}$$

Our objective is to upper bound

$$MSE_{u_0}(Q_h) := \int_0^1 (x - Q_h(x;u_0))^2 dx \tag{4.5}$$

for all u_0.

Lower Bound

We suppose that the initial value u_0 is fixed and known, with $0 \le u_0 < 1$. The unit interval $[0,1]$ is partitioned into subintervals $J = J(\mathbf{q}_x(N))$. The *optimal decoding algorithm*[1] $Q_{opt}(x;u_0)$ maps the quantization data $\mathbf{q}_x(N)$ that are associated with x to the midpoint of the interval $J(\mathbf{q}_x(N))$. There are at most $\frac{(N+1)(N+2)}{2}$ quantization intervals determined by the values given in (4.2). For $u_0 = 1/2$, $\mathcal{S}_N(u_0)$ is the subset of $\mathcal{F}_{2N}$ formed by the fractions with even denominators. Hence, some of the values are repeated, similarly to the case $\mathcal{S}_N(0) = \mathcal{F}_N$ discussed earlier. In this case, the number of distinct values is asymptotic to $\frac{3}{\pi^2}N^2$ as $N \to \infty$, using [7, Theorem 330], since the points in this set can be put in one-to-one correspondence with the Farey sequence $\mathcal{F}_N$. The intervals produced by the Farey sequence $\mathcal{F}_{2N}$ range in size from $\frac{1}{2N}$ down to size $\frac{1}{4N^2}$, and the interval $[0, \frac{1}{2N}]$ contributes $\frac{1}{96}N^{-3}$ all by itself to the MSE of the optimal decoding algorithm. We now show that the same bound holds for an arbitrary u_0.

Theorem 4.1. *For the mean square error of the optimal decoding algorithm, one has the lower bound*

$$MSE_{u_0}(Q_{opt}) \ge \frac{1}{96}N^{-3}, \tag{4.6}$$

for all $u_0 \in [0,1)$.

Proof. We will show that at least one of the open intervals $(0, \frac{1}{2N})$ or $(1 - \frac{1}{2N}, 1)$ contains no quantization threshold. This interval is of length $\frac{1}{2N}$, and since an interval of length $|I|$ contributes $\frac{1}{12}|I|^3$ to the MSE, the contribution of this interval is $\frac{1}{96}N^{-3}$.

- **Case 1:** $0 \le u_0 < 1/2$.

  For $1 \le j \le n$, and $1 \le n \le N$,

  $$\frac{j - u_0}{n} \ge \frac{j - u_0}{N} \ge \frac{1 - u_0}{N} \ge \frac{1}{2N},$$

  hence $(0, \frac{1}{2N})$ contains no quantization threshold.

[1]The optimality of this algorithm is a consequence of the assumption that x is uniformly distributed in $[0,1]$. When conditioned on the data $\mathbf{q}_x(N)$ the distribution of x is uniform on the quantization interval $J(\mathbf{q}_x(N))$.

- **Case 2:** $1/2 \le u_0 < 1$.

  For $1 \le j \le n$, and $1 \le n \le N$,

$$\frac{j-u_0}{n} \le \frac{n-u_0}{n} \le 1 - \frac{u_0}{N} \le 1 - \frac{1}{2N},$$

  hence $(1 - \frac{1}{2N}, 1)$ contains no quantization threshold.

□

The lower bound in Theorem 4.1 is not optimal; the optimal constant seems hard to determine, but we believe it to be about five times larger than $\frac{1}{96}$. We show in [4] the following exact result:

Theorem 4.2. *Suppose that $u_0 = 0$ or $u_0 = \frac{1}{2}$. Then, one has*

$$MSE_{u_0}(Q_{opt}) = \alpha_{u_0} N^{-3} + O(N^{-4} \log N) \quad as \quad N \to \infty, \tag{4.7}$$

where

$$\alpha_0 := \frac{1}{6}\frac{\zeta(2)}{\zeta(3)} = 0.22807,$$

and

$$\alpha_{1/2} := \frac{2}{21}\frac{\zeta(2)}{\zeta(3)} - \frac{1}{16} = 0.06782,$$

where $\zeta(\cdot)$ is the Riemann zeta function.

The proof is based on the explicit relation of the set of quantization thresholds in these two cases with the Farey series. Theorem 4.2 sets a limit on how much the constant $\frac{1}{96}$ in Theorem 4.1 can be improved, since the best constant can be no larger than $\frac{2}{21}\frac{\zeta(2)}{\zeta(3)} - \frac{1}{16}$. Numerical simulations suggest that this bound for $u_0 = 1/2$ is actually close to the minimum over all initial conditions u_0, and conceivably it might give the best constant.

Upper Bound

We view u_0 as fixed with $0 \le u_0 < 1$, but otherwise unknown. The quantization values $(q(1), q(2), \ldots, q(N))$ are known to the decoder. For simplicity we assume that $N = 2M - 1$ is odd. The triangular filter $\{h(n) : 1 \le n \le 2M-1\}$ of mass 1 is given by

$$h(n) = \begin{cases} \frac{1}{M} - \frac{(M-n)}{M^2} & 1 \le n \le M\ , \\ \\ \frac{1}{M} - \frac{(n-M)}{M^2} & M \le n \le 2M-1\ . \end{cases} \tag{4.8}$$

We give a detailed analysis for the case $N = 2M - 1$ only; for the case $N = 2M$ we may discard the value $q(N)$ and use the above filter on the remaining values.

Theorem 4.3. *For the mean square error of the triangular filter decoder, one has the upper bound*

$$MSE_{u_0}(Q_h) \leq \frac{40}{3} N^{-3} \tag{4.9}$$

for all $u_0 \in [0, 1)$.

The proof uses two number-theoretic lemmas, whose statements and proofs are given next. In the following, (m, n) denotes the greatest common divisor of m and n.

Lemma 4.4. *For fixed constant* u_0 *and all positive integers* n *and* m,

$$\left| \int_0^1 \langle nx + u_0 \rangle \langle mx + u_0 \rangle \, dx - \frac{1}{4} \right| \leq \frac{1}{12} \frac{(n, m)}{nm}^2 . \tag{4.10}$$

Proof. We shall prove the lemma by establishing the formula

$$\int_0^1 \langle nx + u_0 \rangle \langle mx + u_0 \rangle \mathrm{d}x = \frac{1}{4} + \frac{1}{12} \frac{(n, m)}{nm}^2 \phi_{n,m} \ , \tag{4.11}$$

where $|\phi_{n,m}| \leq 1$. Denote the expression on the left hand side by $c_{n,m}$. We substitute $nx + u_0$ and $mx + u_0$ for x in the Fourier series expansion

$$\langle x \rangle = \sum_{k=-\infty}^{\infty} a_k e^{2\pi i k x},$$

where $a_k = (-2\pi i k)^{-1}$ for $k \neq 0$ and $a_0 = 1/2$. This Fourier series is only conditionally convergent, and is to be interpreted as the limit as $N \to \infty$ of the sum taken from $-N$ to N. However, its partial sums are uniformly bounded [8, Ex. 4, p. 22]:

$$\left| \sum_{|k| \leq N} a_k e^{2\pi i k x} \right| \leq C, \qquad \text{for all } x \text{ and all } N. \tag{4.12}$$

Hence, using the bounded convergence theorem, one can change the order of integration and double sum to obtain

$$c_{n,m} = \sum_{k \in \mathbf{Z}} \sum_{l \in \mathbf{Z}} a_k \overline{a}_l \ e^{2\pi i (k-l) u_0} \int_0^1 e^{2\pi i (kn - lm) x} \mathrm{d}x. \tag{4.13}$$

Summing up (4.13) over the nonzero indices k, l given by $kn = lm$ and straightforward manipulations result in

$$\begin{aligned} c_{n,m} &= \frac{1}{4} + \frac{1}{4\pi^2} \frac{(n, m)}{nm}^2 \sum_{d \neq 0} \frac{1}{d^2} e^{2\pi i d u_0 (m-n)/(n,m)} && (4.14) \\ &= \frac{1}{4} + \frac{1}{12} \frac{(n, m)}{nm}^2 \phi_{n,m}, && (4.15) \end{aligned}$$

for some $|\phi_{n,m}| \le 1$,[2] upon using $\sum_{d\neq 0} \frac{1}{d^2} = \frac{\pi^2}{3}$. □

Lemma 4.5. *For all positive integers L,*

$$\sum_{n=1}^{L}\sum_{m=1}^{L} \frac{(n,m)}{nm}^2 \le 5L. \tag{4.16}$$

Proof. We have

$$\begin{aligned}
\sum_{n=1}^{L}\sum_{m=1}^{L} \frac{(n,m)}{nm}^2 &= \sum_{d=1}^{L} \sum_{\substack{1\le n,m\le L\\ (n,m)=d}} \frac{(n,m)}{nm}^2 \\
&\le \sum_{d=1}^{L} \sum_{j=1}^{\lfloor L/d\rfloor}\sum_{k=1}^{\lfloor L/d\rfloor} \frac{1}{jk} \\
&= \sum_{d=1}^{L} \left(\sum_{j=1}^{\lfloor L/d\rfloor} \frac{1}{j}\right)^2 \\
&\le \sum_{d=1}^{L} (1+\log\frac{L}{d})^2 \qquad (4.17)
\end{aligned}$$

However, this last expression is bounded by

$$(1+\log L)^2 + \int_1^L (1+\log(L/y))^2 \mathrm{d}y = 5L - 4 - 2\log L,$$

which proves (4.16). [3] □

[2] Using the formula

$$\sum_{d=1}^{\infty} \frac{1}{d^2}\cos d\theta = \frac{1}{4}(\theta-\pi)^2 - \frac{1}{12}\pi^2, \qquad 0\le\theta\le 2\pi,$$

the exact value of $\phi_{n,m}$ is easily found to be

$$\phi_{n,m} = \frac{3}{2}(2\langle u_0 \frac{m-n}{(n,m)}\rangle - 1)^2 - \frac{1}{2}\ .$$

[3] The constant 5 appearing in (4.16) can be improved to $\gamma^2 + 5\gamma/2 + 7/3 \cong 4.1$ by using the inequality

$$\sum_{j=1}^{L} \frac{1}{j} \le \gamma + \log L + \frac{1}{2L}\ ,$$

where γ is the Euler-Mascheroni constant defined by

$$\gamma = \lim_{N\to\infty}\left(\sum_{j=1}^{N}\frac{1}{j} - \log N\right) = 0.5772\ldots$$

Numerical experiments suggest the optimal constant to be ~ 3.

Proof of Theorem 4.3. Suppose $N = 2M-1$ is odd. We set

$$\epsilon_N(x) := x - Q_h(x; u_0)$$

where $Q_h(x; u_0)$ is the triangular filter decoder

$$Q_h(x; u_0) := \sum_{n=1}^{2M-1} h(n)q(n) \ , \tag{4.18}$$

with filter weights (4.8). We have

$$\begin{aligned} \epsilon_N(x) &= \sum_{n=1}^{2M-1} (x - q(n))h_n \\ &= \sum_{n=1}^{2M-1} (u(n) - u(n-1))h(n). \end{aligned}$$

Summing this by parts yields

$$\epsilon_N(x) = -\frac{1}{M^2}(u(0) + ... + u(M-1)) + \frac{1}{M^2}(u(M) + ... + u(2M-1)) \tag{4.19}$$

Then we have

$$|\epsilon_N(x)| \leq \frac{1}{M^2}\left|\sum_{n=0}^{M-1} u(n) - \frac{M}{2}\right| + \frac{1}{M^2}\left|\sum_{n=M}^{2M-1} u(n) - \frac{M}{2}\right|, \tag{4.20}$$

which, upon taking the square yields

$$|\epsilon_N(x)|^2 \leq \frac{2}{M^4}\Big(\sum_{n=0}^{M-1} u(n) - \frac{M}{2}\Big)^2 + \frac{2}{M^4}\Big(\sum_{n=M}^{2M-1} u(n) - \frac{M}{2}\Big)^2. \tag{4.21}$$

We now consider the mean square error,

$$MSE_{u_0}(Q_h) = \int_0^1 |\epsilon_N(x)|^2 \mathrm{d}x.$$

Substituting the value of $u(n)$ into into (4.21), and integrating, we get

$$MSE_{u_0}(Q_h) \leq \frac{2}{M^4}\int_0^1 \left\{ \Big(\sum_{n=0}^{M-1} \langle nx + u_0\rangle - \frac{M}{2}\Big)^2 + \Big(\sum_{n=M}^{2M-1} \langle nx + u_0\rangle - \frac{M}{2}\Big)^2 \right\} \mathrm{d}x. \tag{4.22}$$

We expand this expression, substitute $\int_0^1 \langle nx + u_0\rangle \ \mathrm{d}x = 1/2$ for positive n, and rearrange to get

$$\begin{aligned} MSE_{u_0}(Q_h) \ \leq \ & \frac{2}{M^4}\left\{ (u_0 - \frac{1}{2})^2 + \sum_{n=1}^{M-1}\sum_{m=1}^{M-1}\left(\int_0^1 \langle nx + u_0\rangle\langle mx + u_0\rangle \ \mathrm{d}x - \frac{1}{4}\right)\right. \\ & \left. + \sum_{n=M}^{2M-1}\sum_{m=M}^{2M-1}\left(\int_0^1 \langle nx + u_0\rangle\langle mx + u_0\rangle \ \mathrm{d}x - \frac{1}{4}\right)\right\}. \end{aligned} \tag{4.23}$$

Next we apply Lemma 4.4 to (4.23) and replace the term $(u_0 - \frac{1}{2})^2$ by its maximum value 1/4, to obtain

$$MSE_{u_0}(Q_h) \leq \frac{2}{M^4}\left(\frac{1}{4} + \frac{1}{12}\sum_{n=1}^{M-1}\sum_{m=1}^{M-1}\frac{(n,m)}{nm}^2 + \frac{1}{12}\sum_{n=M}^{2M-1}\sum_{m=M}^{2M-1}\frac{(n,m)}{nm}^2\right). \quad (4.24)$$

Finally, we conclude our estimate of $MSE_{u_0}(Q_h)$ by applying Lemma 4.5 with $L = 2M-1 = N-1$ to (4.24) (combining the double sums) to obtain

$$MSE_{u_0}(Q_h) \leq \frac{40}{3}N^{-3} - \frac{16}{3}N^{-4}, \quad (4.25)$$

which yields the desired bound. ∎

Remark: The proof of Theorem 4.3 did not determine the best constant for MSE using the triangular filter, and some improvements are possible on the constant $\frac{40}{3}$ by more careful argument, since the constant in Lemma 4.5 can be improved slightly.

4.1.2 Approximation in L^p

In Chapters 2 and 3, we have considered pointwise and L^∞ approximation rates of general bandlimited functions using first order $\Sigma\Delta$ quantization. A natural question is whether it is possible to approximate functions in the L^p metric. In this case, we consider functions in the class

$$\mathcal{B}_\pi^p(\mathbb{R},[0,1]) := \{x : \mathbb{R} \to [0,1] : x \in \mathcal{B}_\pi \cap L^p\}, \quad (4.26)$$

for $1 \leq p < \infty$. In the L^∞ setting, the range $[0,1]$ of functions under consideration was arbitrary (however, in agreement with the choice $\{0,1\}$ of quantization levels); by selecting $q(n)$ from $\{a,b\}$ instead, we would be able to handle functions taking their values in $[a,b]$ and essentially everything would be the same. In the L^p setting, however, there is an additional integrability requirement on the functions and this seems to ruin the transposibility. If instead of $\mathcal{B}_\pi^p(\mathbb{R},[0,1])$, we wanted to approximate the class $\mathcal{B}_\pi^p(\mathbb{R},[-1,1]) := \{x : \mathbb{R} \to [-1,1] : x \in \mathcal{B}_\pi \cap L^p\}$ in L^p using functions of the form $\sum_n q(n)\varphi(t - \frac{n}{\lambda})$ with $q(n) \in \{-1,1\}$, this would correspond to a different setting.[4] We shall discuss the setting corresponding to the class (4.26) with $q(n) \in \{0,1\}$ in detail, and then outline the results for the setting with range $[-1,1]$. It turns out that there exists a non-trivial correspondence between the two settings which makes them behave similarly regarding our problem.

We would first like to make an asymptotical analysis of the output bit sequence q_λ produced by the first order $\Sigma\Delta$ quantization algorithm given in (2.24)-(2.26) for functions in $\mathcal{B}_\pi^p(\mathbb{R},[0,1])$. The cases $p = 1$ and $1 < p < \infty$ lead to two different types of behavior, and this becomes important for the approximation problem.

[4]This new setting would be equivalent to approximating functions in $\{x : \mathbb{R} \to [0,1] : x \in \mathcal{B}_\pi$ and $\int |x(t) - \frac{1}{2}|^p dt < \infty\}$ with $q(n)$ drawn from $\{0,1\}$. We would still be measuring the error in the L^p norm, although neither the target function nor the approximant would lie in L^p! (This seemingly odd situation, however, would not cause a problem, since both target and approximant would have the same "dc component" equal to 1/2.)

Case 1: $p = 1$**.**

Let $x \in \mathcal{B}_\pi^1(\mathbb{R}, [0,1])$. The Poisson summation formula[5] gives

$$\frac{1}{\lambda} \sum_n x(\tfrac{n}{\lambda}) = \sum_n \hat{x}(2\pi\lambda n) = \hat{x}(0) \tag{4.27}$$

for all $\lambda > 1$. Since $|\hat{x}(0)| \leq \|x\|_{L^1}$, we get that $\sum_n x(\frac{n}{\lambda}) < \infty$. (In fact, in our case, $\hat{x}(0) = \int x(t)dt = \|x\|_{L^1}$, since $x(t) \geq 0$ for all t.) Clearly, this implies that $|X_\lambda(n)| \leq \lambda \|x\|_{L^1}$ for all n, and thus there is a positive integer n_λ such that $q_\lambda(n) = 0$ for all $|n| \geq n_\lambda$. Without loss of generality, we assume n_λ to be the smallest such number, i.e., $n_\lambda := \min\{n : q_\lambda(m) = 0, \text{ for all } |m| \geq n\}$.

Clearly, the definition of n_λ implies $\sum\limits_{|n|>n_\lambda} x(\frac{n}{\lambda}) < 2$, so that (4.27) yields

$$\lambda \|x\|_{L^1} < 2 + \sum_{|n| \leq n_\lambda} x(\tfrac{n}{\lambda}) \leq 2n_\lambda + 3. \tag{4.28}$$

This gives the lower bound $n_\lambda > \frac{1}{2}(\lambda \|x\|_{L^1} - 3)$. On the other hand, an upper bound may be obtained using

$$u_\lambda(n_\lambda) + \sum_{n > n_\lambda} x(\tfrac{n}{\lambda}) < 1. \tag{4.29}$$

However, it is hard to give a precise estimate because of the variability of $u_\lambda(n_\lambda)$. For a function x that obeys the decay estimate $|x(t)| \leq |t|^{-\beta_0}$, it can at least be said that for every λ, there is an initial condition $u_\lambda(0)$ for which $n_\lambda \leq C\, \lambda^{\beta_0/(\beta_0 - 1)}$. The following theorem states that whenever $n_\lambda = o(\lambda^2)$, one has approximation in L^1.

Theorem 4.6. *Let $x \in \mathcal{B}_\pi^1(\mathbb{R}, [0,1])$, and q_λ be the output of the first order $\Sigma\Delta$ quantizer with input $(x(\frac{n}{\lambda}))$. Assume φ satisfies, together with (2.20), the decay rate $|\varphi(t)| \leq C\,|t|^{-\gamma_0}$ for some $\gamma_0 > 2$. Then*

$$\|x - \tilde{x}_\lambda\|_{L^1} = O\left(\frac{n_\lambda}{\lambda^2}\right) + o(1), \tag{4.30}$$

where $\tilde{x}_\lambda = \frac{1}{\lambda} \sum\limits_{n \in \mathbb{Z}} q_\lambda(n) \varphi(\cdot - \frac{n}{\lambda})$.

Proof. For all $|t| > n_\lambda/\lambda$, the approximant $\tilde{x}_\lambda(t)$ can easily be bounded as

$$\begin{aligned} \left| \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_\lambda(n) \varphi(t - \tfrac{n}{\lambda}) \right| &\leq \frac{1}{\lambda} \sum_{|n| < n_\lambda} |\varphi(t - \tfrac{n}{\lambda})| && (4.31) \\ &\leq C \left| t - \frac{n_\lambda}{\lambda} \right|^{-\gamma_0 + 1}. && (4.32) \end{aligned}$$

[5]Note that the Poisson summation formula is valid since x is necessarily continuous (in fact analytic) and is also of bounded variation (which can be seen from Bernstein's inequality for L^1-bandlimited functions [9, p. 14]).

On the other hand the basic estimate gives $|x(t) - \tilde{x}_\lambda(t)| \leq C/\lambda$ for all t. So, for $T_\lambda > n_\lambda/\lambda$, one has

$$\begin{aligned} \|x - \tilde{x}_\lambda\|_{L^1} &\leq \int_{|t|\leq T_\lambda} |x(t) - \tilde{x}_\lambda(t)|\, dt + \int_{|t|>T_\lambda} (|x(t)| + |\tilde{x}_\lambda(t)|)\, dt \qquad (4.33) \\ &\leq C\frac{T_\lambda}{\lambda} + \int_{|t|>T_\lambda} |x(t)|\, dt + C\left|T_\lambda - \frac{n_\lambda}{\lambda}\right|^{-\gamma_0+2}. \qquad (4.34) \end{aligned}$$

The theorem follows by choosing $T_\lambda = \frac{n_\lambda}{\lambda} + \lambda^\epsilon$ for some $0 < \epsilon < 1$. □

Case 2: $1 < p < \infty$**.**

We would like to consider functions $x \in \mathcal{B}^p_\pi(\mathbb{R}, [0,1])$ with $\sum_{n=1}^\infty x(\frac{n}{\lambda}) = \infty$ for all $\lambda > 1$. Such a function can be constructed very simply. Start with any function y defined on $\hat{\mathbb{R}}$ with the properties

- $\mathrm{supp}(y) \subset [-\frac{\pi}{2}, \frac{\pi}{2}]$,
- $y \in L^{(2p)'} \backslash L^2$, where $(2p)' = \frac{2p}{2p-1}$ is the conjugate index of $2p$ (note that $1 < (2p)' < 2$),
- $\|y\|_{L^1} \leq 2\pi$,
- y is even and real,

and set $x := (y^\vee)^2 = (y * y)^\vee$, where $y^\vee(t) = \frac{1}{2\pi}\int y(\xi) e^{it\xi} d\xi$. Clearly, x is in $\mathcal{B}_\pi$ since $y * y$ is supported on $[-\pi, \pi]$, is everywhere positive since $y^\vee$ is real, is in L^p since $y^\vee \in L^{2p}$. Moreover $\|x\|_{L^\infty} = \|y^\vee\|^2_{L^\infty} \leq (\frac{1}{2\pi}\|y\|_{L^1})^2 \leq 1$. Finally, $x \notin L^1$, for otherwise this would imply $y^\vee \in L^2$, which is impossible since $y \notin L^2$. Thus, $x \in \mathcal{B}^p_\pi(\mathbb{R}, [0,1]) \backslash L^1$. We claim that $\sum_n x(\frac{n}{\lambda}) = \infty$. To see this, define $c^\lambda_n(y)$ to be the n^{th} Fourier coefficient of y, given by $c^\lambda_n(y) := \frac{1}{2\pi\lambda}\int_{-\lambda\pi}^{\lambda\pi} y(\xi) e^{-in\xi/\lambda} d\xi$. Then, $c^\lambda_n(y) = \frac{1}{\lambda} y^\vee(-\frac{n}{\lambda})$, so that $x(\frac{n}{\lambda}) = \lambda^2 (c^\lambda_{-n}(y))^2$. But $y \notin L^2$, so $\sum_n |c^\lambda_n(y)|^2 = \infty$. Note that x is an even function, so that $\sum_{n=1}^\infty x(\frac{n}{\lambda}) = \infty$ as well, proving our claim.

We now look at the asymptotics of $q_\lambda(n)$ as $|n| \to \infty$. First, observe that $x(t) \to 0$ as $|t| \to \infty$ by the Riemann-Lebesgue lemma, since $y * y \in L^1$. This has the following implication: For given positive integer M, let t_M be such that $x(t) < 1/M$ for all $|t| > t_M$. Then, for all $n > \lambda t_M$, at most one among M consecutive output bits $q_\lambda(n+1), \ldots, q_\lambda(n+M)$ can be equal to 1. To see this, just note that

$$\sum_{m=n+1}^{n+M} q_\lambda(m) = -[u_\lambda(n+M) - u_\lambda(n)] + \sum_{m=n+1}^{n+M} x(\tfrac{m}{\lambda}) < 2. \qquad (4.35)$$

A similar statement can be made for negative indices as well. Let $(n^\lambda_k)_{k\in\mathbb{Z}}$ be the increasing sequence of indices at which the output bit is equal to 1. (This is an infinite sequence since $\sum_n x(\frac{n}{\lambda}) = \infty$ implies $|Q_\lambda(n)| \to \infty$ as $|n| \to \infty$.) For a large

positive integer M, let k be such that $n_k^\lambda > \lambda t_M$. Then,

$$2 = \sum_{m=n_k^\lambda}^{n_{k+1}^\lambda} q_\lambda(m) = -[u_\lambda(n_{k+1}^\lambda) - u_\lambda(n_k^\lambda - 1)] + \sum_{m=n_k^\lambda}^{n_{k+1}^\lambda} x(\tfrac{m}{\lambda}) \tag{4.36}$$

$$< 1 + \frac{1}{M}(n_{k+1}^\lambda - n_k^\lambda + 1), \tag{4.37}$$

so that $(n_{k+1}^\lambda - n_k^\lambda) \geq M$. In other words, the sequence $(n_{k+1}^\lambda - n_k^\lambda) \to \infty$ as $k \to \infty$. Similarly, $(n_{k-1}^\lambda - n_k^\lambda) \to -\infty$ as $k \to -\infty$. One can express $\tilde{x}_\lambda$ in terms of the sequence (n_k^λ) as

$$\tilde{x}_\lambda = \frac{1}{\lambda} \sum_{k \in \mathbb{Z}} \varphi(\cdot - \frac{n_k^\lambda}{\lambda}). \tag{4.38}$$

In general this function is not in any L^p space, for $p < \infty$.[6] Thus, one can not expect any convergence in L^p, for $p < \infty$.

Remark: If instead we look at functions in the class $\mathcal{B}_\pi^p(\mathbb{R}, [-1, 1])$ using the first order $\Sigma\Delta$ quantization algorithm with $q(n) \in \{-1, 1\}$, then similar results are still valid. For instance, in the case $p = 1$, the output bit sequence eventually converges to the alternating sequence $(-1)^n$ and it is possible to have convergence in L^1 using slightly more specific filters. On the other hand, for $1 < p < \infty$, we can consider the same subclass of functions $x = (y * y)^\vee$ constructed above. For these functions, the output sequence now behaves asymptotically like $(-1)^n$, except that at each n_k^λ, the phase of the sequence $(-1)^n$ is flipped, i.e., the sequence of indices n_k^λ corresponds to two consecutive $+1$ values. It can again be shown that approximants constructed from such bit sequences in general do not belong to any L^p space.

4.2 Other Schemes, Lower Bounds and Other Information Theoretical Considerations

4.2.1 The Family of Daubechies and DeVore

In Chapter 2, we mentioned higher order $\Sigma\Delta$ schemes and gave the derivation of the basic error estimate for a stable k-th order scheme, borrowed from Daubechies and DeVore [3]. We also mentioned explicit construction of a family of stable schemes for all orders by Daubechies and DeVore; as far as we are aware, this is the first such construction. Below is their scheme and the statement of their stability result. The proof may be found in [3]. It is assumed that the output bits are drawn from $\{-1, 1\}$ instead of $\{0, 1\}$.

Theorem 4.7 (Daubechies-DeVore [3]). *Suppose $|x_n| \leq a < 1$ for all $n \in \mathbb{N}$. For a positive integer k, let $L_1 = \lfloor (5 + 4a)/(1 - a) \rfloor + 2$, $M_1 = 2(1 + a)$, and*

[6] This is trivial to see when φ is compactly supported, since $(n_{k+1}^\lambda - n_k^\lambda) \to \infty$ as $k \to \infty$.

$M_j = (3L_1)^{j-1}4^{(j-1)(j-2)}M_1$ *for* $j = 2, \ldots, k-1$. *Let the sequences* $(q_n)_{n\in\mathbb{N}}$ *and* $(u_n^{(j)})_{n\in\mathbb{N}}$, $j = 1, \ldots, k$ *be defined by*

$$\begin{aligned} u_n^{(1)} &= u_{n-1}^{(1)} + x_n - q_n \\ u_n^{(j)} &= u_{n-1}^{(j)} + u_n^{(j-1)} \qquad j = 2, \ldots, k \\ q_n &= sign\{u_{n-1}^{(1)} + M_1 sign[u_{n-1}^{(2)} + \ldots + M_{k-1} sign(u_{n-1}^{(k)}) \cdots]\}, \end{aligned} \tag{4.39}$$

where $u_{-1}^{(j)} = 0$ *for* $j = 1, \ldots, k$ *as initial conditions and the recursion is started at* $n = 0$. *Then,*

$$|u_n^{(k)}| \leq \frac{1}{2}(3L_1)^{k-1}4^{(k-1)(k-2)}M_1 \tag{4.40}$$

for all n.

For a given order k, this result, combined with Theorem 2.9, implies that it is possible to achieve an error estimate of the form $C_k\lambda^{-k}$, where $C_k \sim c^{k^2}$ for some constant $c > 1$. By selecting a suitable order modulator for each λ, it is possible to beat any power law decay. The choice $k \sim \gamma \log \lambda$ for an appropriate γ would yield an $O(\lambda^{-\beta \log \lambda})$ type decay.

4.2.2 Kolmogorov Entropy and Lower Bounds

Various questions can be formulated regarding the approximation properties of one-bit quantization schemes. The first is: what is the best possible accuracy? A bound can be given using the notion of Kolmogorov ϵ-entropy.[7] Since the space $\mathcal{B}_\pi(\mathbb{R}, [0,1])$ is not compact with respect to the norm $\|\cdot\|_\infty$, one works with restrictions to finite intervals. Let $\mathcal{B}_{\pi,T}(\mathbb{R}, [0,1])$ be the class formed by restrictions of functions $x \in \mathcal{B}_\pi(\mathbb{R}, [0,1])$ to the interval $[-T, T]$. Then the *average entropy per unit length* $\overline{\mathcal{H}}_\epsilon(\mathcal{B}_\pi(\mathbb{R}, [0,1]))$ is defined to be

$$\overline{\mathcal{H}}_\epsilon(\mathcal{B}_\pi(\mathbb{R}, [0,1])) := \lim_{T\to\infty} \frac{1}{2T}\mathcal{H}_\epsilon(\mathcal{B}_{\pi,T}(\mathbb{R}, [0,1]), C[-T, T]). \tag{4.41}$$

It is known due to a result by Kolmogorov and Tikhomirov [20], [21, p. 514] that

$$\overline{\mathcal{H}}_\epsilon(\mathcal{B}_\pi(\mathbb{R}, [0,1])) \sim \log \frac{1}{\epsilon}, \tag{4.42}$$

which implies that for sufficiently large values of T, the minimal cardinality of an ϵ-net for the class $\mathcal{B}_{\pi,T}(\mathbb{R}, [0,1])$ is asymptotic to $(1/\epsilon)^{2T}$. In other words, with λ bits per unit interval, one can construct an ϵ-net for $\mathcal{B}_{\pi,T}(\mathbb{R}, [0,1])$, consisting of $2^{2\lambda T}$ functions, where $\epsilon \sim 2^{-\lambda}$. We have seen that the best performance that has been proved so far for existing $\Sigma\Delta$ families is of the type $\lambda^{-\beta \log \lambda} = e^{-\beta(\log \lambda)^2}$, which falls far short of any exponential accuracy. This inefficiency could be due to different reasons:

[7] The Kolmogorov ϵ-entropy of a compact set A in a metric space X is defined as follows: For a given $\epsilon > 0$, let $N_\epsilon(A)$ be the minimal cardinality of an ϵ-net of the set A in X. That is, there is a discrete subset A_ϵ of X with cardinality $N_\epsilon(A)$ such that the ϵ-neighborhood of A_ϵ contains A. Then, the quantity $\mathcal{H}_\epsilon(A) := \mathcal{H}_\epsilon(A, X) := \log N_\epsilon(A)$ is called the ϵ-entropy of the set A in X.

1. For a given λ, let $\mathcal{C}_{\lambda,T} \subset \{0,1\}^{[-\lambda T, \lambda T]}$ be the collection of all possible bitstreams output by the $\Sigma\Delta$ modulator of optimal order, with inputs from $\mathcal{B}_{\pi,T}(\mathbb{R}, [0,1])$. The cardinality $\#\mathcal{C}_{\lambda,T}$ of this set may be too small to achieve any exponential accuracy. (For instance, a first order $\Sigma\Delta$ quantizer with constant input can produce only $O(N^2)$ distinct bitstreams of length N.)

2. Consider the set of functions

$$\left\{ \frac{1}{\lambda} \sum_n q(n) \varphi(\cdot - \tfrac{n}{\lambda}) : q \in \mathcal{C}_{\lambda,T} \right\} \tag{4.43}$$

for a given reconstruction filter φ. Even if the cardinality of $\mathcal{C}_{\lambda,T}$ is sufficiently large, the restrictions to the interval $[-T, T]$ of functions in the set (4.43) may not have a sufficiently "uniform" spread in $\mathcal{B}_{\pi,T}$ to achieve exponential accuracy. More importantly, for $q \in \mathcal{C}_{\lambda,T}$, let $\mathcal{S}_q$ be the class of functions in $\mathcal{B}_{\pi,T}$ whose output is the bit sequence q. The partition $\{\mathcal{S}_q\}_{q \in \mathcal{C}_{\lambda,T}}$ may itself be "non-uniform". (Note, for instance, that the effective quantizer corresponding to N bits of the output of a first order $\Sigma\Delta$ quantizer with constant input contains intervals that are as long as $1/N$.)

3. Alternatively, even the larger class of functions

$$\left\{ \frac{1}{\lambda} \sum_n q(n) \varphi(\cdot - \tfrac{n}{\lambda}) : q \in \{0,1\}^{[-\lambda T, \lambda T]} \right\} \tag{4.44}$$

may be suffering from the same problem as in item 2.

It is an unsolved problem to determine whether exponential decay of error can be achieved using approximants of the form (4.44). For the constant input case, a step towards this lower bound is taken by Konyagin [22], who showed by means of a purely number theoretical construction that it is possible to achieve an $O(c^{-\sqrt{\lambda}})$ type of error decay. Next, we present this construction which was kindly communicated to us by him.

4.2.3 Konyagin's construction

Theorem 4.8 (Konyagin [22]). *For each $x \in [\frac{1}{3}, \frac{2}{3}]$, and for each $\lambda > 2$, there exists a sequence $q_{\lambda,x} \in \{0,1\}^{\mathbb{Z}}$ such that*

$$\|x - \frac{1}{\lambda} \sum_n q_{\lambda,x}(n) \varphi(\cdot - \tfrac{n}{\lambda})\|_\infty \leq c_1 e^{-c_2 \sqrt{\lambda \log \lambda}}, \tag{4.45}$$

where φ is such that $\hat{\varphi}$ is smooth, supported on $[-2\pi, 2\pi]$, and equal to 1 on $[-\pi, \pi]$.

Proof. For a positive integer, let $S_{a,m} := \{n \in \mathbb{Z} : n \equiv a \pmod{m}\}$. It follows straightforwardly from Poisson's summation formula that if $m \leq \frac{\lambda}{2}$, then

$$\frac{1}{\lambda} \sum_{n \in S_{a,m}} \varphi(t - \tfrac{n}{\lambda}) = \frac{1}{m} \tag{4.46}$$

for all t, and all integers a. The construction of $q_{\lambda,x}$ will be done by finding a disjoint collection of arithmetic progressions $\{S_{a_i,m_i}\}$ such that $m_i \leq \frac{\lambda}{2}$ for all i, and

$$\left| x - \sum_i \frac{1}{m_i} \right| \leq c_1 e^{-c_2\sqrt{\lambda \log \lambda}}. \tag{4.47}$$

The following is the construction of this collection:
1. Given λ (large), pick a prime number l such that $l \asymp \sqrt{\lambda/\log\lambda}$. Let $r \sim l/3$, $r \leq l/3$, and $p_1, p_2, \ldots, p_r$ be the first r primes, where the prime l is excluded. The prime number theorem gives $p_r \sim \frac{l}{3}\log l$.
2. Let $k, k_1, \ldots, k_r$ be integers such that $k_i \leq p_i$ for every $i = 1, \ldots, r$ and $r + k \leq l$, and consider the collection of the following arithmetic progressions:

$$S_{i,j,l,p_i} := \{n \in \mathbb{Z} : n \equiv i \pmod{l},\ n \equiv j \pmod{p_i}\}, \tag{4.48}$$

for $i = 1, \ldots, r$ and $j = 1, \ldots, k_i$, and

$$S_{r+i,l} = \{n \in \mathbb{Z} : n \equiv r + i \pmod{l}\}, \tag{4.49}$$

for $i = 1, \ldots, k$. The Chinese remainder theorem implies that S_{i,j,l,p_i} is an arithmetic progression with difference lp_i and also that any two progressions from the union of the above two collections are disjoint. It is easy to check that $lp_r \leq \lambda/2$. Hence, if $q_{\lambda,x} \in \{0,1\}^{\mathbb{Z}}$ is such that $q_{\lambda,x}(n) = 1$ if and only if n is an element of any of these progressions, then

$$\frac{1}{\lambda}\sum_n q_{\lambda,x}(n) = \sum_{i=1}^{r} \frac{k_i}{lp_i} + \frac{k}{l}. \tag{4.50}$$

3. Any integer A such that

$$\frac{1}{3} \leq \frac{A}{lp_1 \ldots p_r} \leq \frac{2}{3} \tag{4.51}$$

determines $k, k_1, \ldots, k_r$ that satisfy the requirements in item 2 and such that

$$\frac{A}{p_1 \ldots p_r} = k + \sum_{i=1}^{r} \frac{k_i}{p_i}. \tag{4.52}$$

This is shown as follows. First, define $P_i = \prod_{j\neq i} p_j$. Note that $\gcd(P_1, \ldots, P_r) = 1$, so that there exist integers $P_1', \ldots, P_r'$ with $P_1P_1' + \ldots + P_rP_r' = 1$. Let k_i be defined by $k_i = AP_i' \pmod{p_i}$ for each i. Hence,

$$A \equiv \sum_{i=1}^{r} k_i P_i \pmod{p_1 \ldots p_r}. \tag{4.53}$$

Then, $k = (A - \sum_{i=1}^{r} k_i P_i)/(p_1 \ldots p_r)$. Now,

$$\sum_{i=1}^{r} k_i P_i \leq \sum_{i=1}^{r} p_1 \ldots p_r \leq r p_1 \ldots p_r \leq \frac{1}{3} l p_1 \ldots p_r \tag{4.54}$$

so that $k \geq 0$. On the other hand, trivially $k \leq A/(p_1 \dots p_r) \leq 2l/3$. This ensures that $r + k \leq l$.
4. This reduces the problem to choosing the numbers $k, k_1, \dots, k_r$ for a given x. We have just seen that it suffices to choose A while satisfying (4.51). Pick A such that

$$\left| x - \frac{A}{lp_1 \dots p_r} \right| \leq \frac{1}{lp_1 \dots p_r}. \tag{4.55}$$

Now, using the prime number theorem once again, $\log(lp_1 \dots p_r) \sim p_r \asymp l \log l \asymp \sqrt{\lambda \log \lambda}$. This concludes the proof. □

Remarks:
1. Note that the constants are independent of the value of x.
2. A similar scheme for arbitrary bandlimited functions (instead of constant functions) is not known.

4.2.4 Democratic Encoding-Decoding

Despite its non-optimal performance of approximation, $\Sigma\Delta$ quantization is being widely used in applications. One reason for this is the error-robustness of $\Sigma\Delta$ codes that we shall discuss in this section. There is a built-in redundancy in the collection of output bitstreams, enabling one to decode corrupted bitstreams with reasonable errors. For instance, consider N consecutive output bits of the first order $\Sigma\Delta$ modulator with constant input, and assume the value of one of the bits is flipped. Then, for the simple rectangular averaging, the magnitude of the error will be $1/N$, regardless of which bit is flipped. Note that this error is already comparable to the uncertainty of the value of input when all N bits are correct. On the other hand, the situation could be much worse in ordinary binary expansion. If the most significant bit is lost, this would create an uncertainty of magnitude $1/2$. The fact that all bits are treated equally in the decoding of $\Sigma\Delta$ bitstreams is the basis of the "democracy" concept, which was first introduced by Calderbank and Daubechies in [23]. They prove that a "democratic" representation in this sense cannot achieve the optimal exponential accuracy in the case of encoding numbers in $[0, 1]$. However, their definition is not very tight and there are democratic codes which, from other points of view, do not seem to put equal weight on the bits of representation. In this section, we shall make an analysis of democracy with a wider set of definitions.

The abstract problem

Let (X, d) be a compact metric space. By an encoder, we mean a mapping

$$E : X \to \{0, 1\}^{\mathbb{Z}} \tag{4.56}$$

of X to infinite binary sequences[8], or more generally, a family $\{E_I\}_{I \in \Lambda}$ of mappings, where

$$E_I : X \to \{0, 1\}^I \quad \text{for } I \in \Lambda. \tag{4.57}$$

[8]Sometimes, the natural domain of the binary sequences will not be $\mathbb{Z}$, but $\mathbb{N}$.

Here, Λ is a collection of finite intervals in $\mathbb{Z}$ and we call any such E_I a finite encoder. We also require Λ to contain a nested infinite collection of intervals whose union is the whole of $\mathbb{Z}$.

Similarly, a decoder is a mapping

$$D : \{0,1\}^{\mathbb{Z}} \to X, \tag{4.58}$$

or more generally a family $\{D_I\}_{I\in\Lambda}$ of mappings, where

$$D_I : \{0,1\}^I \to X \quad \text{for } I \in \Lambda. \tag{4.59}$$

For a given encoder E_I, it initially suffices to define a decoder only on the range of E_I. However, in order to be able to decode (possibly) corrupted codewords, it may be desirable that the decoder be defined on larger domains.

Let $P_I : \{0,1\}^{\mathbb{Z}} \to \{0,1\}^I$ be the projection operator defined by restriction to the index set I. This way, every encoder $E : X \to \{0,1\}^{\mathbb{Z}}$ gives rise to a family of finite encoders by $E_I = P_I E$. For instance, $\Sigma\Delta$ encoders with constant input are of this type. When only E is given, we will always assume that the finite encoders E_I are obtained in this fashion. Also, when $I = \{1,\dots,N\}$, we shall use the shorthand notations E_N and D_N.

The *accuracy* $\mathcal{A}(E_I, D_I)$ of the encoder-decoder pair (E_I, D_I) is defined to be

$$\mathcal{A}(E_I, D_I) := \sup_{x\in X} d(x, D_I E_I(x)). \tag{4.60}$$

A natural requirement for an encoder-decoder family $\{(E_I, D_I) : I \in \Lambda\}$ is to approximate the points of X with arbitrarily fine accuracy, i.e.,

$$\lim_{|I|\to\infty,\, I\in\Lambda} \mathcal{A}(E_I, D_I) = 0. \tag{4.61}$$

For $x \in X$, its *encoding neighborhood* $V_I(x)$ is defined as

$$V_I(x) := E_I^{-1} E_I(x), \tag{4.62}$$

and for a codeword $b \in Ran\, E_I$, we call $E_I^{-1}(b)$ the *decoding set* of b. Then, to a given encoder E_I, there corresponds an *optimal decoder* $D_{I,opt} : Ran\ E_I \to X$ defined as follows: $D_{I,opt}(b) := \hat{x}$ such that

$$\sup_{y\in E_I^{-1}(b)} d(\hat{x}, y) \le \sup_{y\in E_I^{-1}(b)} d(z, y) \qquad \text{for all } z \in X. \tag{4.63}$$

Such a point always exists (though it may not be unique) and satisfies

$$d(x, \hat{x}) \le \operatorname{diam}(V_I(x)) \qquad \text{for all } x \in E_I^{-1}(b). \tag{4.64}$$

In general, the bound (4.64) cannot be improved. However, if X is a *centered space*,[9] then one can improve on this estimate:

$$d(x, \hat{x}) \le \frac{1}{2} \operatorname{diam}(V_I(x)). \tag{4.65}$$

[9] A centered space is a metric space (X, d) such that for every subset A of X, there is a point in $c \in X$ with the property that $d(x, c) \le \frac{1}{2}\operatorname{diam}(A)$ for all $x \in A$ (see [20]).

Note that with our definitions, it is possible that $D_{I,opt}E_I(x) \notin V_I(x)$, or equivalently, $E_I D_{I,opt}(b) \neq b$. That is, the optimal reconstruction in the sense of (4.63) may not be "consistent". If necessary, one can avoid this inconsistency by forcing $\hat{x}$ to be in $E_I^{-1}(b)$, and replacing $z \in X$ by $z \in E_I^{-1}(b)$ in the minimization condition (4.63). However, it then becomes possible that no minimizer exists.

For a given decoder D_I, it is also possible to define the *optimal encoder* $E_{I,opt}$ via Voronoi regions. That is,

$$E_{I,opt}^{-1}(b) := \{x \in X : d(x, D_I(b)) \leq d(x, D_I(c)) \text{ for all } c \in \{0,1\}^I\}. \tag{4.66}$$

Note that the pair $(E_{I,opt}, D_I)$ is always consistent, i.e., $E_{I,opt}D_I(b) = b$ for all b.

If X is also equipped with a probability measure μ, then one can define a μ-optimum decoder $D_{I,opt}^{\mu}$ by requiring

$$\int_X d(x, D_{I,opt}^{\mu}E_I(x))\ d\mu(x) \leq \int_X d(x, DE_I(x))\ d\mu(x) \tag{4.67}$$

for all $D : Ran\ E_I \to X$. The solution is given by

$$D_{I,opt}^{\mu}(b) = \arg\min_{z \in X} \int_{E_I^{-1}(b)} d(y,z)\ d\mu(y). \tag{4.68}$$

In general $D_{I,opt}^{\mu} \neq D_{I,opt}$. However, for the analogous definition of $E_{I,opt}^{\mu}$ given by

$$\int_X d(x, D_I E_{I,opt}^{\mu}(x))\ d\mu(x) \leq \int_X d(x, D_I E(x))\ d\mu(x) \tag{4.69}$$

for all $E : X \to \{0,1\}^I$, it is always true that $E_{I,opt}^{\mu} = E_{I,opt}$.

Various notions of democracy

The first definition will be that of Calderbank and Daubechies [23]. Their definition is based on the decoder, and is motivated by the following observation [23] about the minimal requirements for the possibility of "equally important bits". For any encoder E_I, one has

$$\sup_{d_H(b,c)=1} d(D_I(b), D_I(c)) \geq \frac{1}{|I|}(\mathrm{diam}(X) - 2\mathcal{A}(E_I, D_I)), \tag{4.70}$$

where d_H denotes the Hamming metric on binary sequences and $|I|$ denotes the cardinality of I. This claim is proved by choosing two points x and y in X with distance $\mathrm{diam}(X)$, and noting that by flipping one bit at most $|I|$ times, one can transform $E_I(x)$ into $E_I(y)$. Since $d(D_I E_I(x), D_I E_I(y)) \geq \mathrm{diam}(X) - 2\mathcal{A}(E_I, D_I)$, at least one of the bit flips should correspond to a jump that is bigger than the right hand side of (4.70). The upper bound for the maximum amount of error that can be made by flipping one bit can therefore, at best, be $C/|I|$. This leads to the following definition:

Definition 4.9 (Calderbank-Daubechies [23]). *A family of maps* $\{D_N : \{0,1\}^N \to [0,1];\ N = 1,2,\ldots\}$ *is called democratic if there exists a constant* C*, independent of* N*, such that*

$$\sup_{d_H(b,c)=1} |D_N(b) - D_N(c)| \leq C/N. \tag{4.71}$$

The following theorem states the impossibility of having democracy and *optimal accuracy* simultaneously.

Theorem 4.10 (Calderbank-Daubechies [23]). *Suppose* $\{D_N : \{0,1\}^N \to [0,1];\ N = 1,2,\ldots\}$ *is an optimally accurate family of maps, i.e., for a constant* C*, independent of* N*, one has*

$$\mathcal{A}(E_{N,opt}, D_N) \leq C2^{-(N+1)}. \tag{4.72}$$

Then this family cannot be democratic.

Let us analyze decoders of the type

$$D_N(b_1,\ldots,b_N) = \sum_{n=1}^{N} w_n b_n, \qquad b_n \in \{0,1\}. \tag{4.73}$$

Here, we allow for different weights $w_n = w_n^N$ for each N, but for convenience we dropped the dependence on N in the notation. For such decoders, it is easy to assess whether they are democratic. Indeed, a decoder of the type (4.73) is democratic if and only if the weights w_n satisfy $|w_n| \leq C/N$ for all $n = 1,\ldots,N$. The two decoders that we analyzed for first order $\Sigma\Delta$ quantization with constant input, namely the rectangular filter decoder and the triangular filter decoder, are thus democratic. On the other hand, the decoder with $w_n = 2^{-n}$, which corresponds to truncated binary expansion, is clearly non-democratic since $w_1 > C/N$. However, it is optimally accurate (with accuracy 2^{-N}). Now, we shall see that it is possible to build democratic decoders $D_N : \{0,1\}^N \to [0,1]$ with accuracy $\mathcal{A}(E_{N,opt}, D_N) \leq C2^{-\gamma N}$. An example for $\gamma = 1/2$ is the following: Let $N = 2M$ and define D_N by

$$w_1 = \ldots = w_{M-1} = \frac{1}{M}, \quad \text{and} \quad w_n = \frac{2^{-(n-M)}}{2M}, \quad \text{for} \quad n = M,\ldots,2M. \tag{4.74}$$

It is easy to check that the accuracy of this pair is bounded by $2^{-M} = 2^{-N/2}$. On the other hand, it is democratic according to Definition 4.9, since $|w_n| \leq 2/N$ for all $n = 1,\ldots,N$. This example can easily be modified to achieve any $\gamma < 1$ by choosing approximately $(1-\gamma)N-1$ of the weights to be equal to $1/(1-\gamma)N$ and the remaining $\gamma N+1$ weights to decrease exponentially from $1/(1-\gamma)N$ to $2^{-\gamma N-1}/(1-\gamma)N$. Despite democracy, these two groups of bits have very different powers of representation. Clearly, Definition 4.9 is too weak to exclude such anomalies; this suggests we look for alternative definitions. The definition of democracy that we give below excludes these possibilities. In order to distinguish this definition from Definition 4.9, we shall use the terms *strong democracy* and *strongly democratic*.

We shall distinguish two problems: First; we have to find a general way of measuring the importance of a given bit that would lead to the definition of a stronger

notion of democracy. Second; we have to decide whether this should be a property of the encoder only, of the decoder only, or of the encoder and the decoder combined.

For the first problem, we start with the same intuitive idea as before: If any of the bits in a codeword is flipped in a democratic encoding scheme, then the error made by decoding the changed codeword should be the same order of magnitude for all locations of the bit flips. (It is clear that one can not have exact equality except for trivial cases.) However, this definition is not complete, for one also has to measure and compare the effect of changing two, three and more bits at a time.

We define the *uncertainty* associated to a codeword to be

$$U_I(b) := \operatorname{diam}(E_I^{-1}(b)). \tag{4.75}$$

In the probabilistic model, an *average uncertainty* may be defined as

$$U_I^{\mu}(b) := \sup_{z \in E_I^{-1}(b)} \frac{1}{\mu(E_I^{-1}(b))} \int_{E_I^{-1}(b)} d(y,z)\, d\mu(y). \tag{4.76}$$

Let the *joint uncertainty* associated with a pair of codewords (b,c) be measured by

$$U_I(b,c) := \operatorname{diam}(E_I^{-1}(b) \cup E_I^{-1}(c)). \tag{4.77}$$

This quantity measures the maximum distance between two points x and y in X such that $E_I(x) = b$ and $E_I(y) = c$. Define for $k = 0, 1, \ldots, |I|$,

$$M_I^{(k)}(b) := \max \{U_I(b,c) \ : \ c \in Ran\ E_I,\ d_H(b,c) = k\} \tag{4.78}$$

and

$$m_I^{(k)}(b) := \min \{U_I(b,c) \ : \ c \in Ran\ E_I,\ d_H(b,c) = k\}. \tag{4.79}$$

In the above definitions, if the set $\{c \in Ran\ E_I : d_H(b,c) = k\}$ is empty, then we set $M_I^{(k)}(b) := m_I^{(k)}(b) := U_I(b)$. Note that $M_I^{(0)}(b) = m_I^{(0)}(b) = U_I(b)$.

Definition 4.11. *A family $\{E_I\}_{I \in \Lambda}$ of encoders is said to be strongly democratic if*

i) $M_I^{(1)}(b) \le c_1\, m_I^{(1)}(b)$, *and*

ii) $M_I^{(k)}(b) \le c_2\, k\, M_I^{(1)}(b)$,

for all $I \in \Lambda$ and for all $b \in Ran\ E_I$, where all constants involved are independent of b and I.

We shall also work with the following weaker definition of democracy:

Definition 4.12. *For $0 < \gamma \le 1$, we say that a family $\{E_I\}_{I \in \Lambda}$ of encoders is γ-strong democratic if it satisfies*

i′) $M_I^{(1)}(b) \le c_1\, [m_I^{(1)}(b)]^{\gamma}$

together with property (ii) in Definition 4.11.

Remarks:
1. Note that 1-strong democracy is strong democracy.
2. The definition of strong (or γ-strong democracy) is entirely based on the encoder. We say that a family $\{D_I\}_{I\in\Lambda}$ of decoders is strongly (or γ-strongly) democratic if the corresponding family $\{E_{I,opt}\}_{I\in\Lambda}$ of optimal encoders has this property.

Now, we shall prove the following "near-characterization" of γ-strong democracy for decoders of the type (4.73):

Theorem 4.13. *Let $\{D_N : \{0,1\}^N \to [0,1];\ N = 1,2,\ldots\}$ be a family of decoders defined by*

$$D_N(b) = \sum_{n=1}^{N} w_N(n) b(n), \tag{4.80}$$

where we assume that all weights are positive, and

$$0 \le 1 - \sum_{n=1}^{N} w_N(n) \le \max_{1\le n\le N} w_N(n) \tag{4.81}$$

for all N. Let w_N^* denote the increasing rearrangement of w_N, i.e.,

$$w_N^*(1) \le w_N^*(2) \le \cdots \le w_N^*(N). \tag{4.82}$$

Then,

a) *If $w_N^*(N) \le c\,[w_N^*(1)]^\gamma$, then the family $\{D_N\}_{N\ge1}$ is γ-strong democratic.*

b) *If the family $\{D_N\}_{N\ge1}$ is γ-strong democratic, then $w_N^*(N) \le c\,[w_N^*(2)]^\gamma$.*

Proof. **a)** Consider the set of points $Q_N := D_N(\{0,1\}^N)$. Since

$$\Big\{ w_N^*(1),\ w_N^*(1) + w_N^*(2),\ \ldots,\ w_N^*(1) + \ldots + w_N^*(N) \Big\} \subset Q_N, \tag{4.83}$$

and because of (4.81), the size of the largest interval defined by $Q_N \cup \{0,1\}$ is bounded by $w_N^*(N)$. This implies that

$$\mathcal{A}(E_{N,opt}, D_N) \le w_N^*(N). \tag{4.84}$$

Given $b \in \{0,1\}^N$, let x and y be two points in $[0,1]$ such that $E_{N,opt}(x) = b$ and $E_{N,opt}(y) = c$, where $d_H(b,c) = 1$. Clearly, $|D_N(b) - D_N(c)| \le w_N^*(N)$. Thus,

$$|x-y| \le |x - D_N(b)| + |D_N(b) - D_N(c)| + |D_N(c) - y| \le 3\, w_N^*(N), \tag{4.85}$$

so that $M_N^{(1)}(b) \le 3\, w_N^*(N)$. On the other hand, it is clear that $m_N^{(1)}(b) \ge |D_N(b) - D_N(c)| \ge w_N^*(1)$. Hence $M_N^{(1)}(b) \le 3c\,[m_N^{(1)}(b)]^\gamma$.

For the second property, note that $M_N^{(1)}(b) \ge w_N^*(N)$, since one can flip the bit with the largest weight. On the other hand, for any codeword c with $d_H(b,c) = k$, one has $|D_N(b) - D_N(c)| \le k\, w_N^*(N)$, so that $M_N^{(k)}(b) \le (k+2)\, w_N^*(N) \le 3k\, w_N^*(N)$. This proves that $\{D_N\}_{N\ge1}$ is γ-strong democratic.

b) We take $b = (0, \ldots, 0)$, and construct c by flipping only the bit that corresponds to the weight $w_N^*(1)$. Then, it is easy to see that $E_{N,opt}^{-1}(0) = [0, \frac{1}{2}w_N^*(1)]$, and $E_{N,opt}^{-1}(c) = [\frac{1}{2}w_N^*(1), \frac{1}{2}w_N^*(1) + \frac{1}{2}w_N^*(2)]$. Hence $m_N^{(1)}(0) = \frac{1}{2}w_N^*(1) + \frac{1}{2}w_N^*(2) \leq w_N^*(2)$. Since $M_N^{(1)}(0) \geq w_N^*(N)$, our assumption that $M_N^{(1)}(0) \leq c\,[m_N^{(1)}(0)]^\gamma$ implies $w_N^*(N) \leq c\,[w_N^*(2)]^\gamma$. □

Remarks:
1. The binary expansion is not γ-strong democratic for any γ, since $w_N^*(N) = \frac{1}{2}$, and $w_N^*(2) = 2^{-N+1}$.
2. The example in (4.74) is not γ-strong democratic for any γ, since $w_N^*(N) = 2N^{-1}$, and $w_N^*(2) = N^{-1}2^{-N/2+1}$.
3. More generally, it is easy to show that for a γ-strong democratic family of decoders as in the above theorem, $w_N^*(N)$ and $w_N^*(2)$ should satisfy the following properties:

$$(N+1)^{-1} \leq \quad w_N^*(N) \quad \leq c\,N^{-\gamma}, \text{ and} \tag{4.86}$$

$$c\,N^{-1/\gamma} \leq \quad w_N^*(2) \quad \leq (N-1)^{-1}. \tag{4.87}$$

Hence, for this family of decoders, 1-strong democracy implies democracy in the sense of Definition 4.9.
4. For symmetric weights, the above theorem turns into an "exact characterization" since then $w_N^*(1) = w_N^*(2)$. An example is the triangular filter decoder defined in (4.8).

Let us show that first order $\Sigma\Delta$ quantization with constant input is not 1-strong democratic as a family of mappings $\{E_N : [0,1] \to \{0,1\}^N;\ N = 1, 2, \ldots\}$. This claim follows from the following observation: Given N, consider the codewords

$$\begin{aligned} b &= (\underbrace{0, \ldots, 0}_{N \text{ times}}, 1, \underbrace{0, \ldots, 0}_{N-1}, 1), \\ c_1 &= (\underbrace{0, \ldots, 0}_{N}, 1, \underbrace{0, \ldots, 0}_{N-1}, 0), \text{ and} \\ c_2 &= (\underbrace{0, \ldots, 0}_{N}, 0, \underbrace{0, \ldots, 0}_{N-1}, 1), \end{aligned}$$

which, from (4.1), are easily seen to be associated to the Farey intervals $E_N^{-1}(b) = \left[\frac{2}{2N+1}, \frac{1}{N}\right)$, $E_N^{-1}(c_1) = \left[\frac{1}{N+1}, \frac{2}{2N+1}\right)$, and $E_N^{-1}(c_2) = \left[\frac{1}{2N+1}, \frac{1}{2N}\right)$. But then

$$U_{2N+1}(b, c_1) = \operatorname{diam}\Big(\Big[\frac{1}{N+1}, \frac{1}{N}\Big)\Big) \asymp \frac{1}{N^2},$$

whereas

$$U_{2N+1}(b, c_2) = \operatorname{diam}\Big(\Big[\frac{1}{2N+1}, \frac{1}{N}\Big)\Big) \asymp \frac{1}{N},$$

although $d_H(b, c_1) = d_H(b, c_2) = 1$. Thus, the property (i) of Definition 4.11 fails to hold. However, if we take instead the weaker condition (i′) for γ-strong democracy with $\gamma = 0.5$, then there is no longer any obstruction. To see this, let x and y be in

$[0,1]$ such that $E_N(x) = b$, $E_N(y) = c$, where $d_H(b,c) = 1$. Let D_N be the rectangular filter decoder. Then

$$|x - y| \leq |x - D_N(b)| + |D_N(b) - D_N(c)| + |D_N(b) - y| \leq 3/N \tag{4.88}$$

so that $M_N^{(1)}(b) \leq 3/N$. On the other hand, we know that $m_N^{(1)}(b) \geq 1/N^2$ since the shortest Farey interval is of this length. Hence, $M_N^{(1)}(b) \leq 3\,[m_N^{(1)}(b)]^{0.5}$. It is much harder to check property (ii) required for strong or γ-strong democracy. A straightforward upper bound for $M_N^{(k)}(b)$ is $3k/N$. However, a lower bound for $M_N^{(1)}(b)$ of the order $1/N$, which would yield property (ii), is not immediate (if it is valid at all). The main problem is the lack of a good understanding of the "small" collection of bitstreams output by the $\Sigma\Delta$ quantizer.

Finally, we would like to give a definition of democracy for a family of encoder-decoder pairs. We assume that the decoders D_I are defined on the whole domain $\{0,1\}^I$. Note that the joint uncertainty for two codewords b, $c \in Ran\ E_I$, as defined in (4.77), means

$$U_I(b,c) = \max_{x \in E_I^{-1}(b)} \max_{y \in E_I^{-1}(c)} d(x,y), \tag{4.89}$$

and is a symmetric function of its arguments. This quantity does not necessarily measure the error made by confusing the codewords b and c, since a decoder is not specified and hence the erroneous reconstruction is not known. When both the encoder and the decoder are given, we define the quantity $\tilde{U}_I(b \mapsto c)$ to measure the maximum error that is made by decoding the codeword c instead of b:

$$\tilde{U}_I(b \mapsto c) := \max_{x \in E_I^{-1}(b)} d(x, D_I(c)). \tag{4.90}$$

Then the analogous definitions for the extremal values of this quantity are:

$$\tilde{M}_I^{(k)}(b) := \max_{d_H(b,c)=k} \tilde{U}_I(b \mapsto c), \tag{4.91}$$

and

$$\tilde{m}_I^{(k)}(b) := \min_{d_H(b,c)=k} \tilde{U}_I(b \mapsto c). \tag{4.92}$$

Definition 4.14. *A family $\{(E_I, D_I)\}_{I \in \Lambda}$ of encoder-decoder pairs is said to be jointly γ-strong democratic if*

i) $\tilde{M}_I^{(1)}(b) \leq c_1\ [\tilde{m}_I^{(1)}(b)]^{\gamma}$, *and*

ii) $\tilde{M}_I^{(k)}(b) \leq c_2\ k\ \tilde{M}_I^{(1)}(b)$,

for all $I \in \Lambda$ and for all $b \in Ran\ E_I$, where all constants involved are independent of b and I.

Remarks:
1. It is easy to see that Theorem 4.13 continues to hold in terms of the joint γ-strong democracy for the family $\{(E_{N,opt}, D_N) : N = 1, 2, \ldots\}$ defined by (4.80).

2. Let E_N be the encoder of the first order $\Sigma\Delta$ quantizer, and $D_{N,rect}$ be the rectangular filter decoder. Let us show that the family $(E_N, D_{N,rect})$ is jointly 0.5-strong democratic. First, it easily follows that

$$\begin{aligned}\tilde{M}_N^{(k)}(b) &\leq \max_{x\in E_N^{-1}(b)} |x - D_{N,rect}(b)| + \max_{d_H(b,c)=k} |D_{N,rect}(b) - D_{N,rect}(c)| \\ &\leq \frac{k+1}{N}\,. \end{aligned} \tag{4.93}$$

On the other hand, clearly $\tilde{m}_N^{(1)}(b) \geq \frac{1}{2N^2}$, since $\tilde{U}_I(b \mapsto c) \geq \frac{1}{2}\mathrm{diam}(E_N^{-1}(b))$ for all c. This proves property (i) for $\gamma = 0.5$. Next, let us see that $\tilde{M}_N^{(1)}(b) \geq 1/N$. This is clear for $b = (0, \ldots, 0)$, and $b = (1, \ldots, 1)$, so assume b has at least an entry 1 and an entry 0. Let the two codewords c_+ and c_- be constructed by flipping any of the 0's to 1 and any of the 1's to 0, respectively. Clearly

$$D_{N,rect}(c_+) - \frac{1}{N} = D_{N,rect}(b) = D_{N,rect}(c_-) + \frac{1}{N},$$

so that for any x, at least one of the two quantities $|x - D_{N,rect}(c_+)|$ and $|x - D_{N,rect}(c_-)|$ is larger than $1/N$. Hence it follows that $\tilde{M}_N^{(1)}(b) \geq 1/N$. This implies property (ii).

Let $D_{N,tri}$ be the triangular filter decoder defined by (4.8), where $N = 2M - 1$. Let us see that the family $(E_N, D_{N,tri})$ is also jointly 0.5-strong democratic. The proof that property (i) holds is the same as in the previous case. For the second property, let us again assume b has at least an entry 1 and an entry 0. Let c_1 be the codeword constructed by flipping the middle bit $b(M)$, and c_2 be the codewords constructed by flipping any of the bits that has the opposite sign of $b(M)$. Since the weight of the middle bit is $1/M$, one has

$$|D_{N,tri}(c_1) - D_{N,tri}(c_2)| \geq \frac{1}{M} + \frac{1}{M^2} \geq \frac{2}{N},$$

so that it is again true that $\tilde{M}_N^{(1)}(b) \geq 1/N$.

4.2.5 Robustness

We saw that individual bits of $\Sigma\Delta$ codes share roles that are distributed more evenly compared to the bits of binary expansion. We analyzed this property within the framework of "democratic" encoding, but one can also view this property as a certain type of "robustness" against bitwise errors. In this respect, the problem of democratic encoding of signals can be viewed as a natural extension of error correcting coding in which the amount of error in reconstruction (in terms of the metric in the signal space) is controlled by the amount of error in the codes (in terms of the Hamming distance). In this section, we shall look at the general problem of "robust encoding" from a different point of view, and in particular, point out robustness of $\Sigma\Delta$ quantization. This point of view requires robustness against other sources of error, which are usually introduced by flawed circuitry in real hardware implementations.

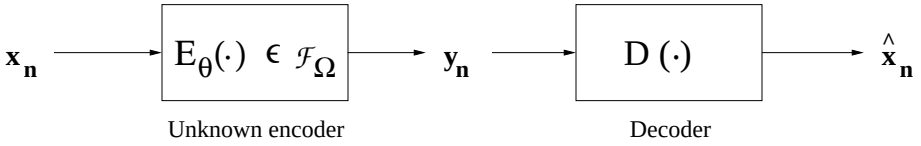

Figure 4.1: Robust encoding problem.

An analog-to-digital conversion scheme is a (necessarily nonlinear) encoding operation that produces an output stream of bits describing a signal. One has an ideal system (the encoder) $E(\cdot)$ used to design the circuit, while the actual hardware circuit can be flawed, resulting in the system $\tilde{E}(\cdot)$ that approximates the behavior of $E(\cdot)$ but which makes certain errors. These errors are most of the time "systematic", but they can occasionally be random as well. Given the same input, the nonlinear nature of the system, combined with feedback, can potentially lead, after some time, to large discrepancies between the internal states of the two systems $E(\cdot)$ and $\tilde{E}(\cdot)$, no matter how small the systematic or random errors are. Hence, the circuit produces extremely "erroneous" bitstreams in Hamming distance. Yet, it is possible that these erroneous codes are "highly correlated" with the correct codes in certain other senses, still leading to close approximations. The robust encoding problem is to design an ideal encoder $E(\cdot)$ for which there exists some decoder robust against certain perturbations of the encoder. Given a robust encoder, the robust decoding problem is to design a decoder $D(\cdot)$ which produces an adequate reconstruction of the signal from the digital bitstreams produced by any of the (not precisely known) systems $\tilde{E}(\cdot)$. Let us note that in contrast to the problem of democratic encoding, the problem of robust encoding is now quite different from error correcting coding.

One can formulate the robust encoding problem as the simple block diagram given in Fig 4.1. We are given a family of encoders $\mathcal{F}_\Omega = \{E_\theta \ : \ \theta \in \Omega\}$ where θ is a (vector) parameter. We may think of these as representing an ideal encoder with the parameter θ measuring the deviation from ideality of a particular hardware implementation. The parameter θ is not under any control, except to lie in a fixed compact set Ω representing the manufacturing tolerances. The decoder's performance is to be measured in the worst case against all the allowable encoders. This model scheme does not include any source of random errors encountered during the execution of the encoding, just systematic encoding errors embodied in the parameter θ.

Consider the quantization step in first order $\Sigma\Delta$ modulation given by (2.11), but instead using a quantizer with *offset* δ. That is,

$$q(n) = Q_\delta(u(n-1) + x(n)), \quad \text{where} \quad Q_\delta(v) = \begin{cases} 1 & \text{if } v \geq 1-\delta, \\ 0 & \text{if } v < 1-\delta. \end{cases} \tag{4.94}$$

Assume also that the initial condition $u_0 := u(0)$ is not necessarily 0. In this case, we can take $\theta = (u_0, \delta)$, but as we show below, the behavior of the system really depends

only on the single parameter $\tilde{\theta} = u_0 + \delta$. The behavior of the system with an ideal quantizer is described by

$$u(n) - u(n-1) = x(n) - Q(u(n-1) + x(n)),\ n = 0, 1, \ldots \tag{4.95}$$

whereas with the non-ideal quantizer it is

$$\tilde{u}(n) - \tilde{u}(n-1) = x(n) - Q_\delta(\tilde{u}(n-1) + x(n)),\ n = 0, 1, \ldots \tag{4.96}$$

The following simple fact, observed in [24], simplifies the robust quantization problem for this case.

Lemma 4.15. *Let x be a fixed input sequence. The output bit sequence q for the non-ideal first order $\Sigma\Delta$ modulator with initial value $\tilde{u}(0) = \tilde{u}_0$ and offset δ is identical to the output bit sequence for the ideal first order $\Sigma\Delta$ modulator with the modified input value $u(0) := \tilde{u}_0 + \delta$.*

Proof. Since $Q_\delta(\cdot) = Q(\cdot + \delta)$, on setting $u(n) := \tilde{u}(n) + \delta$, the system (4.96) becomes equivalent to the system (4.95) with the initial condition $u(0) = \tilde{u}_0 + \delta$. □

Lemma 4.15 shows that studying robustness of a first order $\Sigma\Delta$ modulator against arbitrary initial value $\tilde{u_0}$ and offset error reduces to the special case of studying the ideal system (4.95) with arbitrary (unknown) initial condition u_0. Hence, all of the uniform, pointwise and mean square error estimates that we gave in Chapters 2, 3, and 4 for the first order $\Sigma\Delta$ modulation are robust against offset error in the quantizer, since all these estimates are uniform in the initial condition. Let us note that this reduction is special to first-order $\Sigma\Delta$ modulation. In higher-order schemes the initial value u_0 and offset parameter δ are independent sources of error.

As a comparison, let us look at the effect of quantization offset error in the case of binary expansion of numbers $x \in [0, 1]$. The recursion is via the doubling map on the torus, given by

$$u(n) = \langle 2\, u(n-1) \rangle, \quad n = 1, 2, \ldots \tag{4.97}$$

where $u(0) = x$, and the bits $(b(n))_{n=1}^{\infty}$ are computed by

$$b(n) = Q(2\, u(n-1)), \tag{4.98}$$

where $Q(\cdot)$ is the quantizer $Q_\delta(\cdot)$ defined in (4.94) for $\delta = 0$. Then clearly, one can rewrite (4.97) as

$$u(n) = 2\, u(n-1) - b(n), \tag{4.99}$$

since $b(n) = \lfloor 2\, u(n-1) \rfloor$. Note that the N-term approximation to x, given by $x_N := \sum_{n=1}^{N} 2^{-n} b(n)$, is at most at a distance 2^{-N} from x. Suppose now, that the offset of the quantizer Q_δ is nonzero, producing erroneous bits $\tilde{b}(n)$, and erroneous N-term approximations $\tilde{x}_N := \sum_{n=1}^{N} 2^{-n} \tilde{b}(n)$. It is easy to see that for $1/2 - \delta \leq x \leq 1/2 - \delta/2$, the first bit $\tilde{b}(1)$ is equal to 1, so that $\tilde{x}_N \geq 1/2$, resulting in an error that is bounded below by $\delta/2$. So, the standard decoder is not robust in the asymptotic sense. Let us note that regarding quantizer offset error, robust encoding

and decoding is possible in the case of β-expansions, expansions with respect to a base $\beta < 2$, due to a construction by R. DeVore [25]. The construction achieves its robustness by exploiting the redundancy in representing numbers in such bases. The accuracy of this construction is still exponential in the number of bits N, of the form $2^{-\gamma N}$ for suitable $\gamma > 0$.

Certainly, the possible existence of a robust decoder depends on the family of encoders $\mathcal{F}_\Omega$ considered. For example, in the case of first-order $\Sigma\Delta$ modulation, Feely and Chua [26] consider encoders $\mathcal{F}_\Omega$ that employ *leaky integrators*, meaning

$$u(n) = \beta\, u(n-1) + x(n) - q(n),\ \ n = 0, 1, \ldots \tag{4.100}$$

for some $\beta < 1$. Their results imply that for constant inputs and optimal decoding, the mean square error does not go to zero with increasing the number of output bits, so that a robust decoder does not exist in the asymptotic sense considered here.

Part II

Functional Space Approach to Image Compression

Chapter 5

Introduction

The second part of this thesis is on the "functional space approach" in the mathematical modeling of image compression. Only naturally occuring images are of concern here; compression of cartoons or other man-made graphics is done more efficiently with methods that will not be discussed here. A natural image is a function with a continuous domain and a (bounded) continuous range, where the function's value at a particular point corresponds to the color (or a color component) of that point. However, digital images are discrete valued functions defined on rectangular arrays of picture elements called pixels. The distinction is often suppressed in the mathematical theory, assuming that the discretization, both in space and in amplitude, is sufficiently fine so that the theory and practice resemble each other well. The mathematical theory of the discrete (digital) setting, as well as the mathematical theory of the discrepancy between the continuous and discrete descriptions are quite interesting to study, and involve algebraic elements as well as analysis; concepts like continuity have to be replaced by more cumbersome and less elegant notions, however. We shall not concern ourselves with these aspects here.

Certainly, not every function is an image, at least equally likely; otherwise compression would be impossible. For this reason, understanding the space of images is one of the most important problems of image compression. A typical approach in this direction is to replace the space of images with a simpler mathematical model and study the model instead. When the model fails, one tries to fix it by introducing more complexity. The failure may occur because the simple model space is either too small or too large. Assuming a large class (such as all square integrable functions) has the advantage of guaranteeing that algorithms will always work, though suboptimally. On the other hand, by studying smaller classes (such as the space of piecewise smooth functions) one may devise more efficient algorithms that would work well most of the time but that may sometimes fail badly. The trade-offs are not always clear.

Many image models that have been developed have a stochastic content. In such a model, images are assumed to be the realizations of a random process, where the source of randomness stems from the very way images are viewed as the combinations of several objects each of which has a probabilistic description of location, shape, color and texture. One can add further randomness through environment illumination and perturbative noise introduced by the capturing devices. Ideally, one should then

attempt to derive a probability distribution from such a complex model, but this is too enourmous a task. Instead, one usually performs a leap of faith and assumes one of the standard probability distributions of statistics. Although modeling images via this path is not our purpose here, probabilistic models are occasionally helpful, and we will use them whenever it seems appropriate.

One can also choose to capture the variety and the intrinsic properties of images by using an entirely deterministic model, provided that it is sufficiently rich. The model becomes an appropriate subset $\mathcal{I}$ of a normed space $(X, \|\cdot\|_X)$, where the norm $\|\cdot\|_X$ is used to measure distortion; the set $\mathcal{I}$ itself may be described using much stronger norms, or more complex characterizations. To what extent the norm $\|\cdot\|_X$ approximates the distortion in images as perceived by the human visual system is a separate question that needs to be taken into consideration in practice, but upon which we will not touch. Rather, we will choose to use the standard L^p norms, $1 \leq p \leq \infty$, to measure errors. ([27] has a discussion of which L^p norm in particular is closest to the human visual system's measure of distorion.)

Note that most existing compression algorithms correspond to a "truncated" basis expansion; expansions like these are a classical theme of approximation theory, where the approximant is a linear combination of a finite number of basis functions. In practice, one solves the approximation problem in two steps which are:

1. the choice of a basis, which may well depend on the target function f to approximate, and

2. the selection of a finite subset among those basis functions, together with their coefficients; these will depend on f.

The terms *linear approximation* and *nonlinear approximation* are related to how one makes the choices in 1 and 2 above, and will be explained in Chapter 6. Note that in the above, one may replace the notion "basis" with a much more general family of (not necessarily linearly independent) functions that still span $\mathcal{I}$. We will restrict ourselves to bases here.

From a data compression point of view, the size of the representation is the minimum number of bits needed to encode the approximant. This quantity is called the *rate*. (This point of view requires that the selection of the coefficients must be made from a discrete set so that encoding them requires only finitely many bits. This is referred to as *quantization*.) Rate is a quantity that is not easy to compute, or even to estimate. Instead we shall concentrate on estimating a closely related quantity, the number of parameters of the approximant, to measure the size of the representation; this is customary in approximation theory and also often done in engineering practice as a first indication of the true rate. In a basis expansion, this corresponds to the number of functions employed, without any concern for quantization.

The question of how rate depends on the distortion is the fundamental problem of information theory, whereas how distortion depends on the number of parameters is that of approximation theory. Contrary to the first one, the second problem can often be answered by relating the rate of decay of the approximation error to a "smoothness" condition. In this respect, the classical smoothness spaces in analysis,

such as Hölder, Sobolev and more generally Besov spaces arise naturally in the theory. The correspondence between smoothness and the decay rate of the approximation error is established most completely using wavelet bases. Apart from being part of the most successful practical algorithms for image compression, wavelet bases are also very powerful mathematical tools for precisely the smoothness spaces listed above. A suitable wavelet basis is an unconditional basis for all these Banach spaces and there exist corresponding equivalent discrete norms defined on the wavelet coefficients. It is these norm equivalences that serve as the characterizations of functional spaces in terms of the decay rates of the approximation errors in L^p norms. Additionally, most of the simpler toy models for images are already embedded in various Besov spaces. All of these have formed the basis of the analysis and classification of images in these spaces. This is studied in more detail in Chapter 6.

The purpose of this part of the thesis is to study the appropriateness of these spaces and the methods of modeling image compression in this setting. We find that while this is in general a fruitful approach, it can fail or be misleading in a variety of cases. We discuss these in Chapter 7, together with a more refined approach.

Chapter 6

Nonlinear Approximation and Mathematical Modeling of Image Compression

In this chapter, we present the theoretical foundations of the functional space approach to the mathematical modeling of image compression. Along with these, we review the spaces involved in this approach, which are also fundamental spaces in approximation theory. Then we present the connections between natural images and these function spaces.

6.1 Review of Approximation in Wavelet Bases

6.1.1 Linear and Nonlinear Approximation

Basic Definitions

The problems we shall consider in this section lie in the setting of a Banach space $(X, \|\cdot\|_X)$, in which approximations will take place. Assume that X possesses a (Schauder) basis, and denote it by $\{x_i\}_{i=1}^{\infty}$. The *natural projections* $\{\mathbf{P}_n\}_{n=1}^{\infty}$ associated to $\{x_i\}_{i=1}^{\infty}$ are defined by

$$\mathbf{P}_n\left(\sum_{i=1}^{\infty} a_i x_i\right) = \sum_{i=1}^{n} a_i x_i. \tag{6.1}$$

Denote by X_n the linear span of $\{x_1, \ldots, x_n\}$ (which is also the range of $\mathbf{P}_n$), and consider the problem of finding the best approximation from X_n to a given element f of X. Because the X_n are linear spaces, this defines the setting of *linear approximation*. Define the corresponding approximation error by

$$E_n(f) := \operatorname{dist}(f, X_n)_X = \inf_{g \in X_n} \|f - g\|_X. \tag{6.2}$$

This problem is trivial to solve in the case X is a Hilbert space and $\{x_i\}_{i=1}^{\infty}$ is an orthonormal basis of X, due to Parseval's formula: $\mathbf{P}_n(f)$ is the best linear approx-

imant. In the case of a Banach space, it is also easy if one is interested only in near-best approximants f_n, $1 \le n \le \infty$, defined by

$$\|f - f_n\| \le C\, E_n(f), \tag{6.3}$$

for some constant C. In this case, the natural choice $f_n := \mathbf{P}_n(f)$ would suffice. Indeed, for any $g \in X_n$, $\mathbf{P}_n(g) = g$ and thus

$$\|f - \mathbf{P}_n(f)\|_X \le \|f - g\|_X + \|\mathbf{P}_n(g - f)\|_X, \tag{6.4}$$

which, upon taking the infimum over $g \in X_n$ yields (6.3) with

$$C = 1 + \sup_n \|\mathbf{P}_n\|. \tag{6.5}$$

($\sup_n \|\mathbf{P}_n\|$ is always finite and is called the *basis constant* of $\{x_i\}_{i=1}^{\infty}$.) Similarly, any sequence of uniformly bounded projection operators $\mathbf{Q}_n : X \to X_n$ would result in near-best linear approximation operators.

In contrast to approximations from X_n, an arbitrary *n-term approximation* comes from the nonlinear set Σ_n defined by

$$\Sigma_n := \left\{ \sum_{\lambda \in \Lambda} a_\lambda x_\lambda \ : \ \#\Lambda \le n \right\}; \tag{6.6}$$

hence the name *nonlinear approximation* follows. The error of the best n-term approximation is given similarly by

$$\sigma_n(f) := \inf_{g \in \Sigma_n} \|f - g\|_X. \tag{6.7}$$

When X is a Hilbert space and $\{x_i\}_{i=1}^{\infty}$ is an orthonormal basis of X, the solution of the best nonlinear approximation in the above sense is again easy. Given $f = \sum_i a_i x_i$, it suffices to order $\{a_i\}$ in decreasing magnitude, forming a rearranged sequence $\{a_{\mu(i)}\}$, and set the n-term approximant to be simply $\sum_{i=1}^{n} a_{\mu(i)} x_{\mu(i)}$. However, in arbitrary Banach spaces, even near-best solutions are not easy to find. (The definition of a near-best nonlinear approximant is the same as in (6.3) except that X_n and $E_n(f)$ are replaced by Σ_n and $\sigma_n(f)$, respectively.)

Although the usage of the terms "linear" and "*non*-linear" gives a first impression that these two methods of approximation are mutually exclusive, it is clear from the definitions that nonlinear approximation is truly a superset of linear approximation. The solutions to both problems are trivial in a Hilbert space, and a practician in data compression would immediately consider the nonlinear approximation approach for its generality. The price paid is that given a function f in X, all the coefficients of its expansion in the orthonormal basis $\{x_i\}_{i=1}^{\infty}$ have to be computed to find the n largest contributions. This is a task that can be carried out when X is finite dimensional, as in the case of digitized images.

The generality of nonlinear approximation does not come for free; the necessity of having to specify which basis elements are chosen for a given function introduces an indexing overhead. For arbitrary functions in X, the indices of the selected basis elements (or at least their statistical distributions) can be quite general. When the functions of interest come from a smaller subspace Y of X, there may be more structure in the typical selected index set, and thus it may be thought that a natural ordering of basis elements may lead to near-best solutions of the nonlinear approximation problem. As we will point out later, this assumption is not justified in many cases; nevertheless, this approach is taken in the practical image compression algorithm JPEG, where a particular ordering is employed for the *discrete cosine* basis.

Approximation Spaces

Our aim is not only to find the best (or near-best) approximants f_n for a given element f in X, but also to know the rate of approximation as n goes to infinity. Before we move on to more concrete examples in the following subsections, let us continue with the abstract definition of an *approximation space*.

Suppose (S_n) is an increasing sequence of subspaces of X such that $\bigcup S_n$ is dense in X (as in the case of (X_n) or (Σ_n)). Then, it is natural to consider the class of elements in X that have a given rate of error of approximation from (S_n). For any $\alpha > 0$, let

$$\mathcal{A}^\alpha := \mathcal{A}^\alpha(X, (S_n)) := \{f \in X \,:\, \mathrm{dist}(f, S_n)_X = O(n^{-\alpha})\}. \tag{6.8}$$

Under some technical (but essential) conditions[1] on the spaces (S_n), which are already satisfied by both of the families (X_n) and (Σ_n), it is easy to show that $\mathcal{A}^\alpha$ is a linear subspace of X and the quantity

$$|f|_{\mathcal{A}^\alpha} := \sup_{n\geq 1}\; n^\alpha \,\mathrm{dist}(f, S_n)_X \tag{6.9}$$

defines a semi-norm on $\mathcal{A}^\alpha$.

Although the definition of $\mathcal{A}^\alpha$ has the generality we looked for in the beginning, a finer scale of intermediate spaces $\mathcal{A}^\alpha_q$ is employed for a more complete description. This is done with the addition of the secondary parameter $1 \leq q \leq \infty$ and via the semi-norm defined by

$$|f|_{\mathcal{A}^\alpha_q} := \|(2^{n\alpha}\mathrm{dist}(f, S_{2^n})_X)_{n\geq 0}\|_{l^q}\,. \tag{6.10}$$

(Note that $\mathcal{A}^\alpha_q \subset \mathcal{A}^\beta_p$ if $\alpha > \beta$, regardless of q and p, and $\mathcal{A}^\alpha_q \subset \mathcal{A}^\alpha_p$ if $q < p$. Note also that $\mathcal{A}^\alpha = \mathcal{A}^\alpha_\infty$.) The exact characterizations of these spaces in some concrete settings will be given in the next subsection, when we discuss Besov spaces and wavelet approximations.

[1]Two important conditions are: (i) $aS_n = S_n$ for all $a \neq 0$, (ii) $S_n + S_n \subset S_{cn}$ for some integer constant $c \geq 1$. See [10, pp. 234] for the details.

An important feature of the $\mathcal{A}_q^\alpha$ family is that these spaces can be realized as interpolation spaces between X and a continuously embedded dense subspace $(Y, |\cdot|_Y)$ whenever the so-called *Jackson* and *Bernstein* inequalities are valid for the pair (X, Y) with respect to the family (S_n). A Jackson type inequality (or a *direct estimate*) is said to hold if for some $r > 0$,

$$\operatorname{dist}(f, S_n)_X \leq C\, n^{-r} |f|_Y, \tag{6.11}$$

for all $f \in Y$ and $n \geq 1$. (Note that this holds automatically for the $\mathcal{A}_q^\alpha$ spaces with $r = \alpha$.) A Bernstein inequality (or an *inverse estimate*) controls how fast the semi-norms of elements of S_n grow in Y; it states that

$$|g|_Y \leq C\, n^r \|g\|_X, \tag{6.12}$$

for all $n \geq 1$ and $g \in S_n$. For such a pair (X, Y), it turns out that the real method of interpolation produces precisely the $\mathcal{A}_q^\alpha$ spaces:

$$[X, Y]_{\alpha/r,q} = \mathcal{A}_q^\alpha, \qquad \text{for } \ 0 < \alpha < r, \text{ and all } q. \tag{6.13}$$

(See [10, pp. 235] for a proof of this result.)

6.1.2 Approximation in L^p: Wavelet Bases, Unconditionality and Besov Spaces

In this subsection, we review the approximation theory of functional spaces in wavelet bases, considering both linear and nonlinear approximations. The underlying space X is a space of functions with domain either $\mathbb{R}^d$ or $[0, 1]^d$. We measure errors in L^p norms. Thus, we will be concerned with subspaces of $L^p(\mathbb{R}^d)$ or $L^p([0, 1]^d)$.

A short summary of the main results of multiresolution analysis and wavelet decompositions is given in the Appendix. We set ourselves in the biorthogonal case and employ biorthogonal, compactly supported wavelets with arbitrary smoothness and vanishing moments. We shall establish the Jackson and Bernstein inequalities with respect to the multiresolution analysis generated by the associated scaling function. This will lead to a characterization of the corresponding linear and nonlinear approximation spaces in terms of the classical smoothness spaces. Our exposition will follow DeVore [28] and Cohen [29] closely.

Linear Approximation in Wavelet Bases

Here we study the linear approximation spaces with respect to the family $\{V_j\}_{j \geq 0}$. The first ingredient is an L^p-stability result for the projection operators $\mathbf{P}_{V_j}$, stating

$$\|\mathbf{P}_{V_j} f\|_{L^p} \leq C\, \|f\|_{L^p}, \tag{6.14}$$

uniformly in j. With the same reasoning used in (6.4), this implies that

$$\|f - \mathbf{P}_{V_j} f\|_{L^p} \asymp \operatorname{dist}(f, V_j)_{L^p}. \tag{6.15}$$

(Here the notation $F \asymp G$ denotes equivalance of the quantities F and G: there are two absolute constants C_1 and C_2 such that $C_1 F \leq G \leq C_2 F$, uniformly in all the variables in consideration, unless stated otherwise.) The second ingredient is a result from approximation theory stating that, for any function f in the Sobolev space $W^{r,p}(I)$, where I is a cube of sidelength h, there is a polynomial p of degree at most $r-1$ such that

$$\|f - p\|_{L^p(I)} \leq C\, h^r\, |f|_{W^{r,p}(I)} \tag{6.16}$$

for a constant $C = C(r,p,d)$. The spaces V_j contain polynomials up to degree $r-1$ when the mother wavelet ψ has vanishing moments up to order $r-1$. This, together with the above approximation result, the locality of the basis functions and the L^p-stability of the projectors $\mathbf{P}_{V_j}$ results in the following Jackson inequality:

$$\|f - \mathbf{P}_{V_j} f\|_{L^p} \leq C\, 2^{-jr}\, |f|_{W^{r,p}}. \tag{6.17}$$

The corresponding Bernstein inequality, on the other hand, is merely a consequence of the locality and the L^p-stability of the decompositions.[2] With $\varphi \in W^{r,p}$, the inverse estimate is

$$|f|_{W^{r,p}} \leq C\, 2^{jr}\, \|f\|_{L^p}, \tag{6.18}$$

for all $f \in V_j$.

It remains to apply the characterization result stated in §6.1.1 for $\mathcal{A}_q^\alpha$ spaces. By setting $X = L^p$, $Y = W^{r,p}$, and $S_{2^j} = V_j$, it follows that

$$\mathcal{A}_q^\alpha(L^p, (S_n)) = [L^p, W^{r,p}]_{\alpha/r,q}. \tag{6.19}$$

This result characterizes the approximation spaces with respect to the family $\{V_j\}_{j\geq 0}$ in terms of interpolation spaces between some of the classical spaces in analysis. It is also a classical result in interpolation theory that $[L^p, W^{r,p}]_{\alpha/r,q} = B^\alpha_{p,q}$, namely the Besov space with α order of smoothness in L^p with "fine adjustment parameter" q. Hence, (6.19) turns into the following characterization:

$$\{f \in L^p : (2^{j\alpha} \mathrm{dist}(f, V_j)_{L^p})_{j\geq 0} \in l^q\} = B^\alpha_{p,q}. \tag{6.20}$$

Besov spaces have originally been defined in terms of L^p-moduli of smoothness, as we will describe in the next section when we discuss other properties of these spaces relevant to natural images. Let us mention also that by using the discrete Hardy inequalities (see, e.g. [10]), it follows that

$$\|f\|_{L^p} + \|(2^{j\alpha}\|f - \mathbf{P}_{V_j} f\|_{L^p})_{j\geq 0}\|_{l^q} \asymp \|f\|_{L^p} + \|(2^{j\alpha}\|\mathbf{P}_{W_j} f\|_{L^p})_{j\geq 0}\|_{l^q}, \tag{6.21}$$

which yields a characterization of $\|f\|_{B^\alpha_{p,q}}$ in terms of the wavelet coefficients: If

$$f = \sum_{\lambda\in\Gamma_0} c_\lambda \varphi_\lambda + \sum_{j\geq 0} \sum_{\lambda\in\Lambda_j} d_\lambda \psi_\lambda, \tag{6.22}$$

[2] For any $f = \sum c_\lambda \varphi_\lambda \in V_j$, this means $\|f\|_{L^p} \asymp \|(c_\lambda)\|_{l^p}$, assuming φ_λ are normalized in L^p.

is the wavelet decomposition of f, then

$$\|f\|_{B^\alpha_{p,q}} \asymp \|f\|_{\mathcal{A}^\alpha_q(L^p)} \tag{6.23}$$

$$\asymp \|(c_\lambda)_{\lambda\in\Gamma_0}\|_{l^p} + \|(2^{j\alpha}\|(d_\lambda)_{\lambda\in\Lambda_j}\|_{l^p})_{j\geq 0}\|_{l^q}, \tag{6.24}$$

where we have assumed that the wavelets and the scaling functions were normalized in L^p. If an L^2-normalization is employed, then (6.24) becomes

$$\|f\|_{B^\alpha_{p,q}} \asymp \|(c_\lambda)_{\lambda\in\Gamma_0}\|_{l^p} + \|(2^{j\alpha}2^{jd(1/2-1/p)}\|(d_\lambda)_{\lambda\in\Lambda_j}\|_{l^p})_{j\geq 0}\|_{l^q}. \tag{6.25}$$

This norm equivalence has another important implication: it states that the wavelet bases that we considered are *unconditional bases* for these Besov spaces. An unconditional basis in a Banach space X is a Schauder basis $\{x_i\}$ such that for every $x \in X$, the series expansion $x = \sum_i a_i x_i$ converges unconditionally, that is, $\sum_i a_{\sigma(i)} x_{\sigma(i)}$ converges for every permutation σ of the indices. This is also equivalent to the convergence of $\sum_i \epsilon_i a_i x_i$ for every choice of signs $\epsilon_i = \pm 1$. The unconditionality of wavelet bases just follows from the fact that the right hand side of the norm equivalence (6.24) is a function of the absolute values of the coefficients only.

Nonlinear Approximation in Wavelet Bases

The above norm equivalences also serve to characterize the nonlinear approximation spaces in wavelet bases. Perhaps the most important case is approximation in L^2. We first derive the Jackson and Bernstein inequalities for this case. We also assume that the wavelet basis is orthonormal.

The specific Besov spaces $B^r_{\tau,\tau}$ where $1/\tau = 1/2 + r/d$ arise naturally in this setting. For this index combination, the norm equivalence (6.25) reduces to

$$\|f\|_{B^r_{\tau,\tau}} \asymp \|(c_\lambda)_{\lambda\in\Lambda}\|_{l^\tau}, \tag{6.26}$$

where Λ denotes the set of indices of all the coefficients in the expansion.

As we pointed out in §6.1.1, the best $n-term$ nonlinear approximant $f_n \in \Sigma_n$ in the L^2-norm is given by

$$f_n = \sum_{\gamma\in\Upsilon_n} c_\gamma \psi_\gamma = \sum_{k=1}^{n} c_{\gamma_k}\psi_{\gamma_k}, \tag{6.27}$$

where $\Upsilon_n = \{\gamma_1, \ldots, \gamma_n\}$ is defined to be the set of the indices of the n largest coefficients. The sequence $(c_\lambda)_{\lambda\in\Lambda}$ is in l^τ, hence in weak-l^τ.[3] Thus,

$$k^{1/\tau}|c_{\gamma_k}| = \#\{\lambda : |c_\lambda| \geq |c_{\gamma_k}|\}^{1/\tau}|c_{\gamma_k}| \tag{6.28}$$

$$\leq \|(c_\lambda)_{\lambda\in\Lambda}\|_{l^\tau}; \tag{6.29}$$

[3]The weak-l^q space is defined to be the space of all sequences (x_n) satisfying:

$$\sup_{M>0} \#\{n : |x_n| \geq M\}^{1/q} M < \infty.$$

The space l^q is trivially embedded in weak-l^q, with the l^q norm bounding the above quantity.

which implies, together with (6.26), the estimate

$$|c_{\gamma_k}| \le C\, k^{-1/\tau} \|f\|_{B^r_{\tau,\tau}}. \tag{6.30}$$

Thus, the accuracy of n-term nonlinear approximation is estimated by

$$\begin{aligned} \|f - f_n\|_{L^2} &= \left(\sum_{k=n+1}^{\infty} |c_{\gamma_k}|^2 \right)^{1/2} && (6.31) \\ &\le C\, n^{-1/\tau+1/2} \|f\|_{B^r_{\tau,\tau}}, && (6.32) \end{aligned}$$

resulting in the Jackson inequality with respect to the family $\{\Sigma_n\}$:

$$\mathrm{dist}(f, \Sigma_n)_{L^2} \le C\, n^{-r/d} \|f\|_{B^r_{\tau,\tau}}. \tag{6.33}$$

The inverse estimate is also easily obtained from (6.26). Let $f = \sum_{\lambda\in\Lambda} c_\lambda \psi_\lambda$ be in Σ_n. Then, since there are only n non-zero elements in $(c_\lambda)_{\lambda\in\Lambda}$, we have $\|(c_\lambda)\|_{l^\tau} \le n^{1/\tau-1/2} \|(c_\lambda)\|_{l^2}$, by Hölder's inequality. Hence, it follows that

$$\|f\|_{B^r_{\tau,\tau}} \le C\, n^{r/d} \|f\|_{L^2} \quad \text{for all } f \in \Sigma_n. \tag{6.34}$$

As a result, the nonlinear approximation spaces $\mathcal{A}^\alpha_q(L^2, (\Sigma_n))$ are given by

$$\mathcal{A}^\alpha_q(L^2, (\Sigma_n)) = [L^2, B^r_{\tau,\tau}]_{\alpha d/r, q}. \tag{6.35}$$

The norm equivalence (6.26) indicates that the space $B^r_{\tau,\tau}$ for $1/\tau = 1/2 + r/d$ is isometric to l^τ. Hence, the interpolation space given in the right hand side of (6.35) is just $[l^2, l^\tau]_{\alpha d/r, q} = l^{s,q}$, a Lorentz space. Here, $\frac{1}{s} = \frac{1}{2}(1 - \alpha d/r) + \frac{1}{\tau}(\alpha d/r)$. If we restrict ourselves to the case $q = s$, then $l^{s,s} = l^s \sim B^{\alpha d}_{s,s}$, since then the relation $1/s = 1/2 + \alpha$ holds. Hence, the approximation space is again a Besov space and (6.35) turns into

$$\{f \in L^2 : (2^{n\alpha} \mathrm{dist}(f, \Sigma_{2^n}))_{n\ge 0} \in l^s, 1/s = 1/2 + \alpha\} = B^{\alpha d}_{s,s}. \tag{6.36}$$

Characterization of the nonlinear approximation spaces in L^p is somewhat harder. Interestingly, it turns out that (see [30]), a near-best n-term nonlinear approximant for $f = \sum_{\lambda\in\Lambda} c_\lambda \psi_\lambda$ can be found by selecting the indices with the n largest contributions of $\|c_\lambda \psi_\lambda\|_{L^p}$. The approximation spaces are then given by

$$\mathcal{A}^\alpha_q(L^p, (\Sigma_n)) = [L^p, B^r_{\tau,\tau}]_{\alpha d/r, q}. \tag{6.37}$$

Similar to the L^2 case, there is a particular value of q (satisfying $1/q = 1/p + \alpha$) for which the interpolation space on the right hand side of (6.37) is a Besov space.

Let us now make an efficiency comparison of linear and nonlinear approximation. Here, by efficiency, we understand the amount of smoothness required for a given rate of approximation. For simplicity, let us place ourselves in the setting of $X = L^2([0,1]^d)$. In this case, a linear approximant from V_j employs $N = \dim(V_j) \asymp 2^{jd}$

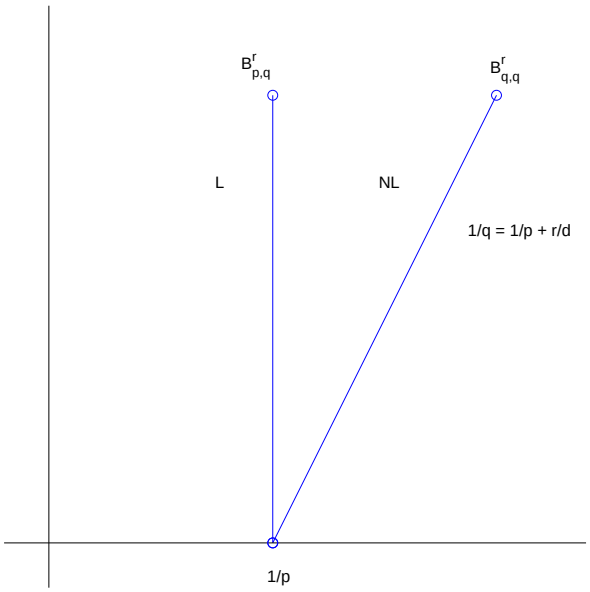

Figure 6.1: Graphical interpretation of linear and nonlinear approximation.

basis functions. Then, (6.20) says that for the linear approximant to achieve an accuracy of $N^{-\alpha/d}$, the target function should (roughly[4]) have α derivatives in L^2. On the other hand, (6.35) states that an N-term nonlinear approximant achieves the same accuracy when the target function has α derivatives in L^s, where $1/s = 1/2 + \alpha/d$, which is a weaker condition. The situation is usually illustrated with the functional space plot of Figure 6.1, in which spaces correspond to points in the the plane. The generic condition of "r derivatives in L^p" (including the spaces $W^{r,p}$ and $B^r_{p,q}$ with arbitrary q) is represented by the point with the coordinates $(1/p, r)$. Then, the two lines denoted by L and NL mark the spaces characterized by the same rate of linear and nonlinear approximation, respectively.

6.2 Embedding Images into Function Spaces

In this section, we look more closely at the various function spaces that have occured as approximation spaces in the previous section and study other properties of these spaces that make them relevant in modeling natural images. We also consider the space BV of functions of bounded variation and a stochastic function space proposed by Cohen and d'Ales as a toy model.

[4]Here, we shall ignore the third parameter of Besov spaces, which is insignificant for the purpose of discussion.

6.2.1 Piecewise Smooth Functions

The whole motivation of this model is the existence of edges in images, which separate objects of different colors. By smooth, we mean C^α for some $\alpha > 1$. More precisely, consider a partition $\Omega_1 \cup \ldots \cup \Omega_M$ of the unit square into domains with smooth (say C^2) boundaries and finite perimeters such that the image $f : [0,1]^2 \to [0,1]$ is C^α on every Ω_i.

Note that this model immediately neglects the texture in an image, which would not fit in a C^α setting. Similarly, noisy images are not considered, either. Images that contain objects with very jaggy boundaries are also excluded from this model.

Two approaches can be taken in the analysis of the rate of approximation in this space. The first approach considers the smallest Besov space that contains the class of piecewise smooth functions with the given regularity and then directly borrows the approximation result for this space. We will do this in §6.2.2. A second approach is more direct, which involves estimating the size of the wavelet coefficients at each level and position, depending on whether the support of the wavelet hits an edge or not. Let us sketch this second analysis:

Let the support of the compactly supported mother wavelet ψ be I. Then the support $I_{j,k}$ of a given wavelet $\psi_{j,k}$ is the set $2^{-j}(I + k)$. If $I_{j,k}$ intersects an edge, let us call this a "type-I" coefficient. For the size $|c_{j,k}|$ of a type-I coefficient, one cannot hope to get an estimate that is better than the trivial estimate:

$$|c_{j,k}| = |\langle f, \psi_{j,k} \rangle| \leq \|f\|_{L^\infty} \|\psi_{j,k}\|_{L^1} \leq C\, 2^{-j}. \tag{6.38}$$

For any level j, there can be at most $O(2^j)$ type-I coefficients, since the $2^{-j}\mathrm{diam}(I)$-neighborhood of the set of domain boundaries contains only this many dyadic squares at this level. Thus, the total L^2 contribution of these coefficients is bounded by $C\, 2^{-j/2}$.

Each of the remaining wavelets will be supported fully in the interior of some domain Ω_i. These define the "type-II" coefficients. We assume that ψ has vanishing moments up to order $\lfloor \alpha \rfloor$. On each $I_{j,k}$, consider a polynomial P of degree $\leq \lfloor \alpha \rfloor$ such that $|f(x) - P(x)| \leq \|f\|_{C^\alpha} |x - x_0|^\alpha$ for some $x_0 \in I_{j,k}$. Since $\langle \psi_{j,k}, P \rangle = 0$, the size $|c_{j,k}|$ of a type-II coefficient is easily estimated by

$$|c_{j,k}| \leq \|(f - P)\chi_{I_{j,k}}\|_{L^\infty} \|\psi_{j,k}\|_{L^1} \leq C\, 2^{-j(\alpha+1)}. \tag{6.39}$$

Hence, at level j, these coefficients contribute at most $2^{-j\alpha}$ to the L^2 norm. Then it follows that

$$\|f - \mathbf{P}_{V_j} f\|_{L^2} \leq C \left(\sum_{l > j} (2^{-l} + 2^{-2l\alpha}) \right)^{1/2} \leq C\, 2^{-j/2}, \tag{6.40}$$

where we have assumed that $\alpha \geq 1/2$. This is the error decay of linear approximation. Noting that 2^{2j} coefficients have been used in this representation, it follows that the linear approximation error $E_N(f)$ decays as $N^{-1/4}$ in the number of basis functions employed in the representation.

Estimates (6.38) and (6.39) also serve to bound the nonlinear approximation error through a weak-type estimate as follows: Given an $\epsilon > 0$ threshold, let N_ϵ be the number of coefficients of absolute value larger than ϵ. Hence,

$$\begin{aligned} N_\epsilon &\leq \#\{(j,k) : \ c_{j,k} \text{ is type-I and } \ \epsilon \leq C\, 2^{-j}\} \\ &\quad + \#\{(j,k) : \ c_{j,k} \text{ is type-II and } \ \epsilon \leq C\, 2^{-j(\alpha+1)}\} \\ &\leq \sum_{2^l \leq C\epsilon^{-1}} C\, 2^l + \sum_{2^{l(\alpha+1)} \leq C\epsilon^{-1}} 2^{2l} \\ &\leq C\, (\epsilon^{-1} + \epsilon^{-2/(\alpha+1)}). \end{aligned} \tag{6.41}$$

which is dominated by $C\,\epsilon^{-1}$ if $\alpha \geq 1$. On the other hand,

$$\begin{aligned} \sigma_{N_\epsilon}(f) &\leq C \left(\sum_{2^l > C\epsilon^{-1}} 2^{-l} + \sum_{2^{l(\alpha+1)} > C\epsilon^{-1}} 2^{-2l\alpha} \right)^{1/2} \\ &\leq C\, (\epsilon + \epsilon^{2\alpha/(\alpha+1)})^{1/2} \\ &\leq C\, N_\epsilon^{-1/2}. \end{aligned} \tag{6.42}$$

This shows that nonlinear approximation is superior to linear approximation, improving the exponent by 1/4. Clearly, this was due to the sparse distribution of the large coefficients in the wavelet expansion. Note that the smoothness of f did not really matter much, as long as α was greater than 1. This is really the nature of the things with this space; it is possible to improve the exponent 1/2 to 1 by more general approximation schemes, such as adaptive triangulations of the domains Ω_i.[5] Note also that this kind of limitation for the rate of approximation in a wavelet basis is typical for *two* dimensions. In one dimension, the situation is much different. We can repeat the same calculation we did above, however with the fundamental difference that the number of type-I coefficients is now bounded by a uniform constant C for all levels. It would then follow that $E_n(f)$ is still limited (however to $O(n^{-1/2})$ this time), whereas $\sigma_n(f)$ can be bounded by $O(n^{-\alpha})$. Hence, for one dimensional piecewise smooth target functions, the performances of linear and nonlinear approximation diverge further apart.

6.2.2 Besov Spaces and BV

There are a number of equivalent ways of defining Besov spaces. The original definition is via moduli of smoothness; however, equivalent characterizations by means of interpolation theory, Littlewood-Paley decompositions, or approximation spaces also exist. We already borrowed some of these characterizations without really worrying about the original definition of Besov spaces.

In terms of their definitions, Besov spaces are closest to the generalized Lipschitz spaces $\mathrm{Lip}^*(\alpha, L^p)$, but with an extra parameter q to refine these spaces further. Let

[5] But not more, based on the Kolmogorov entropy of the unit ball of this space. Recently, Candès and Donoho showed that their *curvelet* expansion gets very close this asymptotic lower bound, losing only a logarithmic factor [31].

us recall the definition of the L^p modulus of smoothness of a multivariate function defined on a domain $\Omega \subset \mathbb{R}^d$. Let Δ_h be the forward difference operator with step $h \in \mathbb{R}^d$, i.e., $\Delta_h f(x) = f(x+h) - f(x)$ and Δ_h^r be the r^{th} power of this operator for integers $r \geq 1$.[6] The r^{th} order L^p modulus of smoothness of a function f is defined by

$$\omega_r(f,t)_p := \sup_{|h| \leq t} \|\Delta_h^r f\|_{L^p}. \tag{6.43}$$

Note that for $r = 1$ and $p = \infty$, this definition reduces to the standard definition of modulus of continuity, except that in addition, one also assumes f is uniformly continuous. Let $r = \lfloor \alpha \rfloor + 1$. Then, the space $\mathrm{Lip}^*(\alpha, L^p)$ is defined to be all functions $f \in L^p$ satistying $\omega_r(f,t)_p \leq Mt^\alpha$, and with its seminorm being the smallest such M. In similarity with this definition, the Besov space $B^\alpha_{p,q}$ is defined to be the space of all functions $f \in L^p$ such that

$$|f|_{B^\alpha_{p,q}} := \left(\int_0^\infty [t^{-\alpha} \omega_r(f,t)_p]^q \frac{dt}{t} \right)^{1/q} < \infty. \tag{6.44}$$

The Besov norm is given by $\|f\|_{L^p} + |f|_{B^\alpha_{p,q}}$. By definition, $B^\alpha_{p,\infty} := \mathrm{Lip}^*(\alpha, L^p)$. Let us note that when α is not an integer, $B^\alpha_{p,p}$ turns out to be identical to the fractional Sobolev space $W^{\alpha,p}$.

Besov spaces may easily contain discontinuous functions even for large values of the smoothness parameter α. A classical example is a piecewise smooth function. Consider a function f defined on $[-1,1]$ having a single jump discontinuity at the origin and being C^α otherwise. Then for small t, $\omega_r(f,t)_p$ can be estimated by $(|O(t^\alpha)|^p + O(r)t)^{1/p} = O(t^{\min(\alpha,1/p)})$. This implies that $f \in B^\alpha_{1/\alpha,\infty}$ and $f \in B^{\alpha'}_{p,q}$ for all $\alpha' < \min(\alpha, 1/p)$ and all q; that is, f preserves its α order of smoothness provided that the smoothness is measured in a sufficiently large L^p space. (Besov spaces are defined by the same expression also for $p < 1$, except that they are only quasi-normed spaces for this range.) In particular, f is in $B^{\alpha'}_{\tau,\tau}$, with $1/\tau = \alpha' + 1/2$, for all $\alpha' < \alpha$, implying that $\sup\{\alpha' : \sigma_n(f)_{L^2} = O(n^{-\alpha'})\} = \alpha$. While these arguments do not immediately say that $\sigma_n(f)_{L^2} = O(n^{-\alpha})$, we know that this error estimate holds for the class of piecewise smooth functions, as we discussed in §6.2.1. Let us also note that the space $\mathcal{A}^\alpha_\infty(L^2, (\Sigma_n))$, which is characterized by the property $\sigma_n(f)_{L^2} = O(n^{-\alpha})$, fails to be a Besov space.

We had seen in §6.2.1 that in two dimension, things were not as impressive. Let f be piecewise smooth on a domain in $\mathbb{R}^2$. Then, one still has $\omega_r(f,t)_p = O(t^{\min(\alpha,1/p)})$, however, for f to belong to the Besov space $B^\mu_{\tau,\tau}$ with $1/\tau = \mu/2 + 1/2$, the condition $\tau < 1/\mu$ immediately implies that $\mu < 1$. Hence, it follows that $\sigma_n(f)_{L^2}$ can only be bounded by $O(n^{-1/2})$, and not better, again verifying the result that was found in §6.2.1 directly by arguments on the size of the coefficients.

Another important space in connection with these classes is the space BV of functions of bounded variation. Apart from its classical definition on the real line, the space BV has a number of different but coinciding definitions on a general domain

[6] A technical assumption regarding bounded domains is the following: $\Delta_h^r f(x) := 0$ whenever any of $x, x+h, \ldots, x+rh$ is not in Ω.

in $\mathbb{R}^d$. We are interested in two of these definitions. The first definition identifies BV with the space $\mathrm{Lip}(1, L^1)$, that is, the class of functions whose L^1 moduli of continuity $\omega_1(f,t)_1$ behave as $O(t)$. It immediately follows from this definition that piecewise smooth functions belong to BV. On the other hand, $\mathrm{Lip}(1, L^1)$ does not belong to the family of Besov spaces. However, the following holds:

$$B^1_{1,1} \subset \mathrm{BV} = \mathrm{Lip}(1, L^1) \subset \mathrm{Lip}^*(1, L^1) = B^1_{1,\infty}. \tag{6.45}$$

This already leads to comparative error decays for linear and nonlinear approximation of functions in BV. In the special case of $\mathrm{BV}([0,1]^2)$, it is shown in [32] that actually $\mathrm{BV} \subset B^{1,w}_{1,1}$, where the latter space is by definition characterized by requiring wavelet coefficients to be in weak-l^1. (For $B^1_{1,1}$, wavelet coefficients are precisely in l^1, as we stated in 6.1.2.) This result is stronger than (6.45).

The second (and more familiar) definition sets BV to be the class of functions in L^1 whose distributional derivatives are Radon measures. (It is known that these two definitions are equivalent for a wide class of domains $\Omega \subset \mathbb{R}^d$.) The relevance of this definition to images may be explained using the co-area formula for BV functions [33]. The co-area formula, when stated for Lipschitz functions, is the following: If $f : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz mapping, then

$$\int_{\mathbb{R}^d} |Df|\, dx = \int_{-\infty}^{\infty} \mathcal{H}^{d-1}(\{x : f(x) = t\})\, dt, \tag{6.46}$$

where $\mathcal{H}^{d-1}$ denotes the $d-1$ dimensional Hausdorff measure on $\mathbb{R}^d$. For BV functions, a similar result holds with the modifications that Df is now a Radon measure and thus the left hand side is replaced by $|f|_{BV}$, and $\{x : f(x) = t\}$ is replaced by the (measure theoretic) boundary of $\{x : f(x) > t\}$ (see [33] for a precise statement of this.) Thus, it follows that a BV image will have contours of finite length. In particular, for the piecewise smooth model, the edges must have finite lengths. Quoting [33, pp. 209], a BV function is "measure theoretically piecewise continuous" with "jumps along a measure theoretically C^1 surface (curve for $d = 2$)". This approximation can be used to justify a BV model for images. However, due to the above corollary of the co-area formula, this model would exclude the possibility of fractal objects in an image such as clouds, which are occasionally modeled as having fractal boundaries. We leave further discussion of this to the next chapter.

6.2.3 Stochastic Setting: Model of Cohen and d'Ales

As we mentioned in the Introduction, it is customary to view images as realizations of a random process. Many models have been proposed in the image processing literature to model the statistics of image pixel intensities and the expansion coefficients in wavelet as well as trigonometric bases. In contrast to these traditional approaches, a stochastic function space model was proposed recently by Cohen and d'Ales in [34]. The model is a one dimensional process $f(t)$ defined on $[0,1]$, and is described by the following ingredients: $0 = d_0 < d_1 < \ldots < d_L < d_{L+1} = 1$ are a random number of points obtained by a Poisson process (of intensity μ), at which discontinuities are to

be placed. On each interval $[d_i, d_{i+1}]$, $f(t)$ is the realization of a zero mean stationary process $X(t)$ whose autocorrelation function $r_X(t)$ is assumed to be C^α for $\alpha > 3/2$. The following are proved in their paper regarding the performances of linear and nonlinear approximation in trigonometric and wavelet bases:

- For linear approximation, the performances of both the trigonometric basis and a sufficiently regular wavelet basis are equivalent to the optimal performance given by the Karhunen-Loeve basis. All of these methods result in

$$\mathbf{E}[E_n(f)_{L^2}^2] \asymp n^{-1}. \tag{6.47}$$

  The poor performance of the Karhunen-Loeve basis follows mainly as a consequence of the low regularity of the global autocorrelation function $r_f(t) = e^{-\mu|t|} r_X(t)$.

- On the other hand, the wavelet basis outperforms the trigonometric basis by a large margin in nonlinear approximation. For the trigonometric basis, one still has $\mathbf{E}[\sigma_n(f)_{L^2}^2] \asymp n^{-1}$, whereas nonlinear approximation in a wavelet basis results in an $O(n^{-\alpha})$ bound for this quantity.

In this section, we would like to point out a few facts regarding the almost sure regularity implied by this model; we shall provide the deterministic function spaces in which almost every realization of this process lies. In particular, this leads to estimates for the error decay rate that hold almost surely.

We consider the stationary stochastic process $X(t)$ defined on $[0, 1]$ whose autocorrelation function $r_X(t)$ is C^α. Then, in particular, there exists a polynomial of degree $\lfloor \alpha \rfloor$ such that

$$|r_X(t) - P(t)| \leq C|t|^\alpha \tag{6.48}$$

for all t. It follows that, for any integer $k > \alpha$,

$$\begin{aligned} \mathbf{E}[|\Delta_h^k X(t)|^2] &= \sum_{i=0}^{k}\sum_{j=0}^{k}(-1)^{i+j}\binom{k}{i}\binom{k}{j}\mathbf{E}[X(t+ih)X(t+jh)] \\ &= \sum_{i=0}^{k}\sum_{j=0}^{k}(-1)^{i+j}\binom{k}{i}\binom{k}{j}\left(r_X(ih-jh) - P(ih-jh)\right) \\ &\leq C|h|^\alpha \end{aligned} \tag{6.49}$$

for all t. We have here made use of the fact that $\sum_{j=0}^{k}(-1)^j\binom{k}{j}P(ih - jh) = \Delta_{-h}^k P(ih) = 0$ for every i. By integrating (6.49) with respect to t, and applying Fubini's theorem, we get

$$\mathbf{E}[\|\Delta_h^k X(\cdot)\|_{L^2}^2] \leq C|h|^\alpha. \tag{6.50}$$

Yet another application of Fubini leads to

$$\mathbf{E}\left[\int_{-1}^{1}[|h|^{-\theta}\|\Delta_h^k X(\cdot)\|_{L^2}]^2\frac{dh}{|h|}\right] < \infty \tag{6.51}$$

for every $\theta < \alpha/2$. At this point we note the following equivalence for Besov norms

$$\int_{-1}^{1} [|h|^{-\theta} \|\Delta_h^k X(\cdot)\|_{L^2}]^2 \frac{dh}{|h|} \asymp \int_{0}^{1} [h^{-\theta} \omega_k(X,h)_2]^2 \frac{dh}{h} \,, \tag{6.52}$$

which is a consequence of $\omega_r(f,t)_p \asymp \frac{1}{t}\int_0^t \|\Delta_s^r f\|_{L^p} ds$ (where the latter quantity is called an averaged modulus of smoothness), and Hardy's inequalities [10]. Thus,

$$\mathbf{E}[\|X\|_{B^\theta_{2,2}}^2] < \infty \tag{6.53}$$

for every $\theta < \alpha/2$. Certainly, this implies that $\|X\|_{B^\theta_{2,2}} < \infty$ except maybe for a measure zero event $\mathcal{E}_\theta$. Choose any sequence $\{\theta_i\}_{i=1}^\infty$ that converges to $\alpha/2$ from below. It is true that $X \in \bigcap B^{\theta_i}_{2,2}$, except maybe for the event $\bigcup \mathcal{E}_{\theta_i}$, which is still measure zero. Because of the nestedness of Besov spaces in the smoothness index, it follows that

$$X \in \bigcap_{\theta < \frac{\alpha}{2}} B^\theta_{2,2} \quad \text{almost surely.} \tag{6.54}$$

Let us note that our approach to derive (6.53) was direct in the sense that we used Besov norms only in their original definitions. However, by estimating the expected value of the squares of the wavelet coefficients ($\mathbf{E}[|c_{j,k}|^2] \leq C\, 2^{-(\alpha+1)j}$ as given in [34]), and using the wavelet characterization of Besov norms, the same result can also be derived indirectly.

We now know that almost surely, the realizations of the piecewise smooth model of Cohen and d'Ales are functions that are piecewise $B^\theta_{2,2}$ for any $\theta < \alpha/2$. The error decay rate of the wavelet expansion for such a function can be estimated similar to the way it was done in §6.2.1 and §6.2.2 for piecewise C^α functions. For instance, it easily follows that for $p \leq 2$, one has $\omega_r(f,t)_p = O(t^{\min(\theta,1/p)})$. Thus, $\sigma_n(f)_{L^2} = O(n^{-\theta})$, which means that almost surely, realizations can be approximated at a rate arbitrary close to the rate of the expected error.

Chapter 7

Studying Images in Besov Spaces: How reliable?

In Chapter 6 we showed, using various characterization results for linear and nonlinear approximation by wavelet bases, that smoothness spaces, in particular the family of Besov spaces, form a potentially convenient setting for studying the accuracy of image compression algorithms and for modeling and classifying natural images; this approach was originally proposed in [27]. In this chapter, we shall carry out a "case study" to test the reliability of this approach. We will start with a discussion of measuring the (Besov) smoothness of images, and then analyze some instances in which this analysis leads to wrong, misleading, or unreliable conclusions. Next we describe an experiment illustrating another source of mismatch that stems from the *ortho-symmetry* of Besov spaces in wavelet bases. Connected with this aspect of our discussion, we prove a theorem on the asymptotic sign structure of wavelet coefficients for the piecewise smooth image model. Finally, we will present a more refined analysis of smoothness through the multifractal formalism and some of its consequences for images.

7.1 Measuring the Smoothness of an Image

We are primarily interested in measuring the Besov smoothness of images. More formally, given an image f, we would like to know the set of smoothness parameters (α, p, q) for which $f \in B^{\alpha}_{p,q}$. Often, one restricts oneself to the "worst" value of q, computing, for each $p > 0$, the smoothness index $\alpha_f(p)$, which we define by

$$\alpha_f(p) := \sup\{\alpha : f \in B^{\alpha}_{p,\infty}\}. \tag{7.1}$$

It was proposed in [27] to estimate Besov smoothness of images via the decay rate of approximation errors in wavelet bases. The norm equivalences that were stated in Chapter 6 serve as the main tool for this purpose. Let us first consider the equivalence given by (6.25). For computational purposes, it is reasonable at the first glimpse to interpret this as

$$\|(d_\lambda)_{\lambda\in\Lambda_j}\|_{l^p} \sim C\, 2^{-j\alpha_f(p)} 2^{-jd(1/2-1/p)}, \tag{7.2}$$

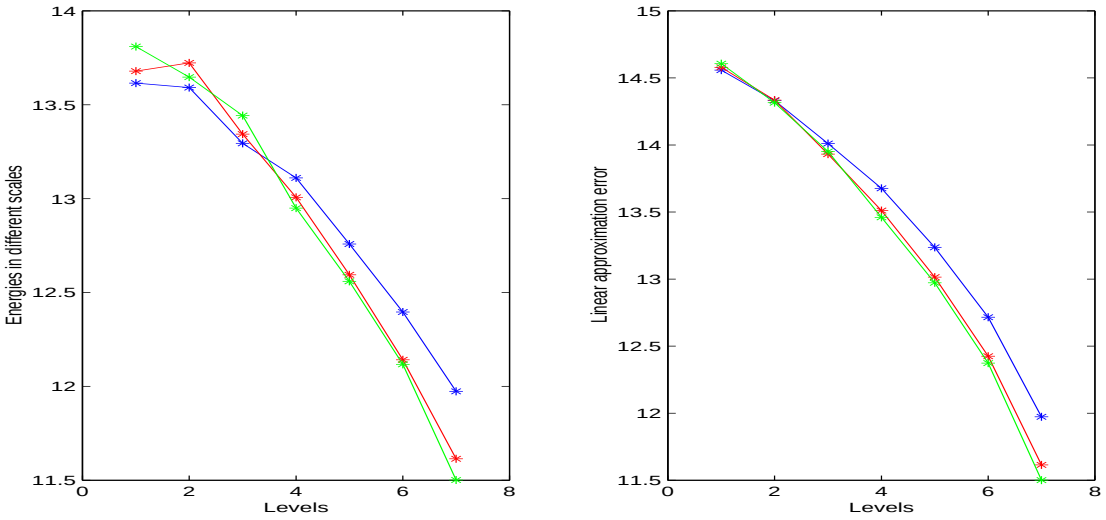

Figure 7.1: Linear approximation error for the Lena image. The logarithm of the energy in each scale and the approximation error at that scale is plotted.

where we have assumed an L^2-normalization for the wavelets and the scaling functions in the decomposition

$$f = \sum_{\lambda \in \Gamma_0} c_\lambda \varphi_\lambda + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} d_\lambda \psi_\lambda. \tag{7.3}$$

This suggests to estimate $\alpha_f(p)$ by $m_j - m_{j+1} - d(1/2 - 1/p)$, where we define

$$m_j := \log_2(\|(d_\lambda)_{\lambda \in \Lambda_j}\|_{l^p}). \tag{7.4}$$

In practice, a typical digitized image contains no more than 6 to 8 dyadic scales available for numerical computation and it is usually hard to observe a steady exponential decay of the approximation error across these scales. We shall give examples illustrating this situation in the next section. Among the most popularly used reference images, the numerically best-behaved example is the Lena image. We plot in Figure 7.1 the graphs of two quantities: On the left is $\log_2(\|(d_\lambda)_{\lambda \in \Lambda_j}\|_{l^2})$, $j = 1, 2, \ldots, 7$, for the first three wavelets in the Daubechies family (which have 1, 2, and 3 vanishing moments, respectively). On the right is the quantity $\log_2(\mathrm{dist}(f, V_j)_{L^2})$ for $j = 1, 2, \ldots, 7$. If $\|(d_\lambda)_{\lambda \in \Lambda_j}\|_{l^2}$ had a truly exponential decay (which would appear as a linear behaviour in the graph), the same would hold for the second quantity as well. Based on the slopes of these graphs where the behaviour is close to linear, it may be argued that $\alpha_f(2)$ is around 0.4. That is, the image Lena has "0.4 order of smoothness in L^2".

Let us note that in these computations, wavelet coefficients have been computed by applying the pyramidal algorithm to the pixel values directly. This implicitly assumes, in fact, that

$$f = \sum_{\lambda \in \Gamma_J} p_\lambda \varphi_\lambda, \tag{7.5}$$

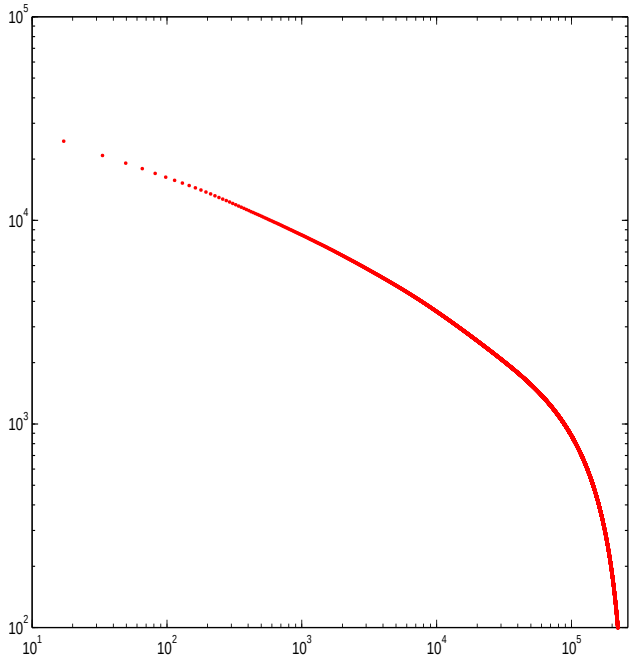

Figure 7.2: Log-log plot of $\sigma_n(f)$ in L^2 for the Lena image.

where $(p_\lambda)_{\lambda\in\Gamma_J}$ is the array of pixels in the image - an assumption which is not correct but nevertheless reasonably approximate.

Let us next consider the characterization (6.36) of nonlinear approximation in L^2. In this case, one looks for a power-law for the decay of approximation error $(\sigma_n(f))_{n=1}^N$, where N is the number of pixels in the image. Following a similar approximate argument as in the linear approximation case leads to

$$\sigma_n(f) \sim C\, n^{-\alpha_f(s)/2}, \quad \text{where} \quad \frac{1}{s} = \frac{1}{2} + \alpha. \tag{7.6}$$

Figure 7.2 is a log-log plot of $\sigma_n(f)$ in L^2 for the Lena image, where the wavelet has two vanishing moments. The slope is approximately 0.3 so that $s \approx (0.5+0.3)^{-1} = 5/4$, and hence $\alpha_f(5/4) \approx 0.6$. That is, the image has "0.6 order of smoothness in $L^{5/4}$".

We have thus identified two Besov spaces in which the Lena image lies minimally. Note that none of these spaces is embedded into the other through a Sobolev-type embedding. (This can be seen by checking the index combinations in the characterizations, or perhaps noting in Figure 6.1 that all Sobolev-type embeddings occur on lines parallel to the line marked as "NL".) Consequently, there are two different ways of estimating the function $\alpha_f(\cdot)$: for each p, one can use the characterization of linear approximation or nonlinear approximation in L^p. In Figure 7.3, we marked with a circle the two spaces that we had found earlier. We also computed $\alpha_f(p)$ for a range of p using the formula (7.2), i.e., within the framework of linear approximation, and plotted the result as a dotted curve in the same figure. This curve is our numerical

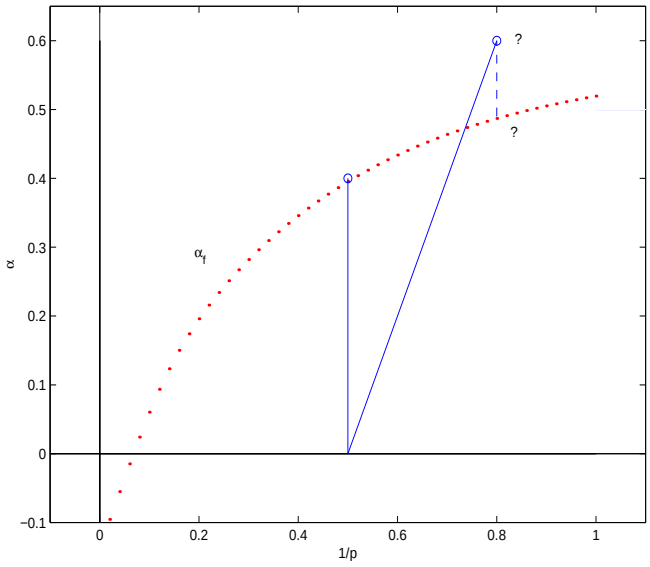

Figure 7.3: An estimate of $\alpha_f(\cdot)$ for the Lena image using the characterization of linear approximation. The two points marked by question marks correspond to two different estimates using nonlinear approximation in L^2 and linear approximation in $L^{5/4}$.

estimate of the best possible smoothness index as a function of $1/p$. We could also have used nonlinear approximation (i.e., apply formula (6.37)) to estimate this curve numerically; this is harder and except for the one point ($1/p = 0.8$, $\alpha_f(p) = 0.6$) which we found earlier, we did not carry out these computations. Note that this point lies some distance from the smoothness curve estimated by linear approximation, illustrating the inaccurateness of our estimates.

7.2 Problems Caused by Ambiguity in the Measurements

At what scale does the asymptotic regime start?

There is not always a definite slope: In the previous section, we worked with the Lena image to illustrate how smoothness is "measured" by inspecting the asymptotical behavior of linear and nonlinear approximation error in wavelet bases. Although there were only a limited number of scales in the Lena image, it was still possible to observe consistent exponential and inverse polynomial decays. However, this is only occasionally the case. We tested this procedure on various images with different characteristics and there seemed to be no generally consistent decay profile. One of the major problems was the lack of a definite slope in the logarithmic plots. We have picked the Window image (see Figure 7.9) to illustrate this example. Figures 7.4 and 7.5 plot the same graphs of Figures 7.1 and 7.2, but for the Window image instead

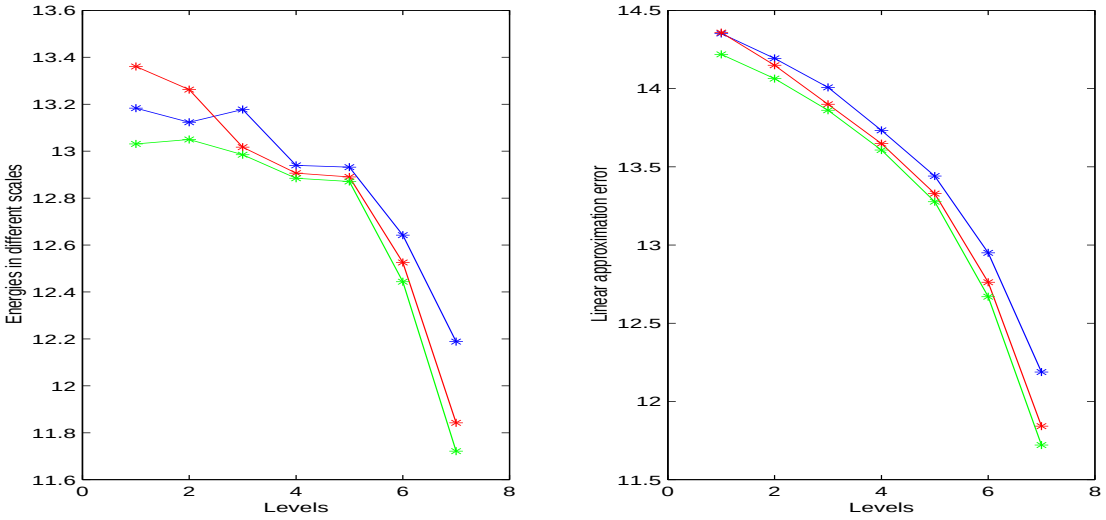

Figure 7.4: Linear approximation error for the Window image. The logarithm of the energy in each scale and the approximation error at that scale is plotted.

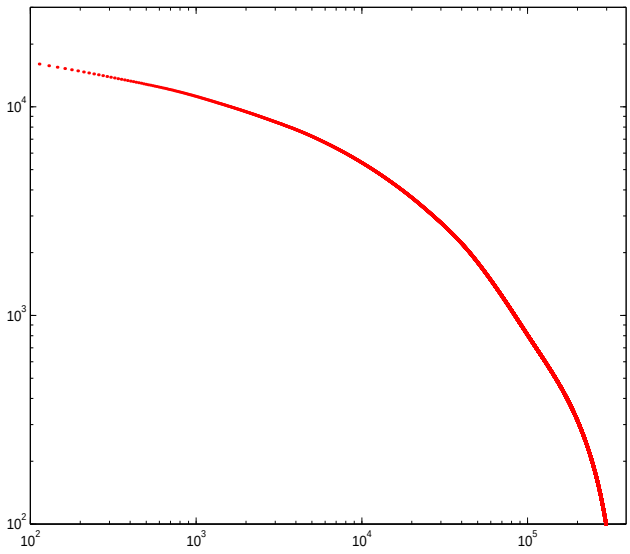

Figure 7.5: Log-log plot of $\sigma_n(f)$ in L^2 for the Window image.

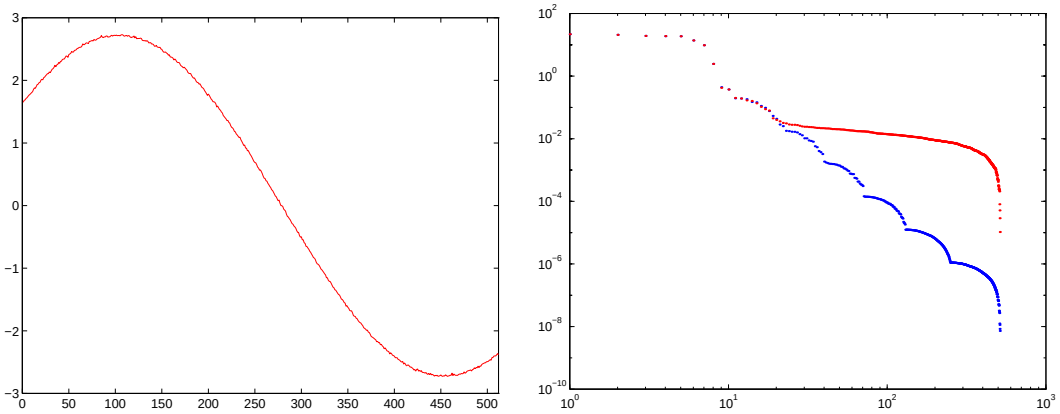

Figure 7.6: Noisy sinusoidal signal on the left and wavelet coefficients in decreasing order of magnitude corresponding to the noiseless and noisy signals.

of Lena. As it can be seen, asymptotics can start "too late" or may not at all!

Noise

One reason to study the smoothness of images is to understand how smooth a noise-free image really is. The term "noise-free" is not well defined; sometimes, irregular behavior that resembles noise in an image may very well be due to texture. In this next example case, we study a situation that leads to misleading conclusions about the intrinsic smoothness of images. We carry out a one dimensional simple experiment in which we look at $f = f_s + f_n$, where f_s and f_n are the smooth and noisy components. The smooth component f_s is assumed to be C^α for a sufficiently large α (determined by the number of vanishing moments of the wavelet used in the expansion). In order to carry out the numerical experiment, we select a sufficiently fine sampling resolution for f_s and draw each noise sample from a Gaussian random variable of variance $\sigma \ll \|f\|_\infty$. In Figure 7.6, we plot a realization of this experiment for which f_s is a pure sinusoid. The noisy signal is on the left. On the right, we have the log-log graphs of the wavelet coefficients of f_s and f, ordered in decreasing magnitude. The fast decaying curve corresponds to f_s and the slope of its envelope is determined by the number of vanishing moments of the mother wavelet. The curve that lies above it corresponds to f. It can be seen that the asymptotics of the noise has kicked in at a very early stage, and no indication of the smoothness of f_s can be read from this plot, although the actual data on the left is still very "smooth" when considered as a scan line of an image. We thus conclude that the decay can be very slow in the presence of noise, even if the noise level is very small.

Mixtures

Consider a piecewise smooth function f whose "components" have different degrees of smoothness; i.e., f is C^{α_i} (and not better) on the domain Ω_i, and not all α_i are the

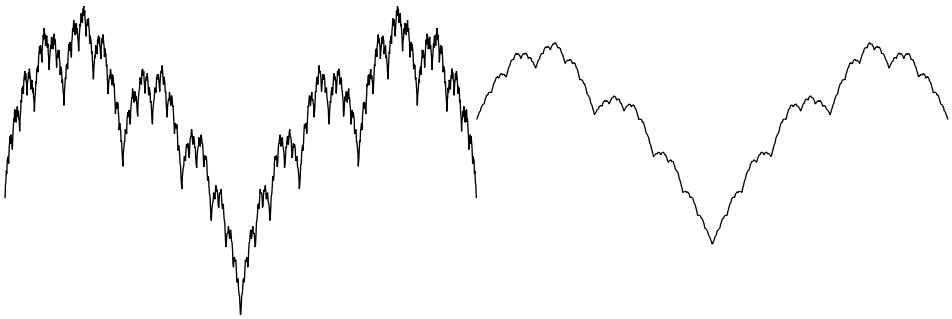

Figure 7.7: A function which is composed of two pieces with different Hölder smoothness.

same. An example would be an image with textured regions. The smoothness of texture would be very small, whereas object with smooth color shades would correspond to a much higher Hölder index. We know from the theory presented in the previous chapter that the asymptotical behavior of the sorted coefficients will be determined by the edges and/or the region of smallest Hölder index. (In one dimension, there would not be any contribution of the singular points.) However, the finite range of indices in a digitized image may easily distort the asymptotical behavior in such a mixture case. As a toy example to illustrate this, let us consider a one dimensional signal composed of two components with **exact** C^{α_1} and C^{α_2} Hölder smoothness.[1] We plot in Figure 7.7 such a function that was created using lacunary Fourier series. We have sampled this function and expanded in a suitable wavelet basis and plotted the coefficients in decreasing order of magnitude in Figure 7.8. If we had all the scales in the sampled data, the coefficient decay would be determined by the region of smaller smoothness index, but in the finite setting, we eventually run out of the big coefficients, and start seeing coefficients of the more regular region. However, the curve is then shifted, and the slope of the tail in the log-log plot does not give the actual smoothness index of that region. One can approximate the situation as in the following example: Suppose we would like to sort the union of two finite sequences $\{n^{-\alpha_1}\}_{n=1}^{N}$ and $\{n^{-\alpha_2}\}_{n=1}^{N}$, where $\alpha_1 < \alpha_2$. Let this new sequence be c_n. For $n > N + M$, where $M := N^{\alpha_1/\alpha_2}$, one has $c_n = (n - N)^{-\alpha_2}$. However, on a log-log plot, the slope

$$\frac{\log c_{n+1} - \log c_n}{\log(n+1) - \log n} \tag{7.7}$$

of this portion of the curve would no longer be constant, and it changes from $-\alpha_2(1 + N/M)$ to $-2\alpha_2$ on the interval $[N + M, 2N]$. Even the minimum value of the slope, being $2\alpha_2$, is twice what one would expect. This is why the measurement becomes misleading: there is no way to accurately estimate the smoothness by looking at the asymptotic decay. Note that this is independent of the size of the data, and sampling at a finer density would not improve the situation!

[1] This can easily be achieved using Fourier series of the type $\sum 2^{-n\alpha} \cos 2^n x$.

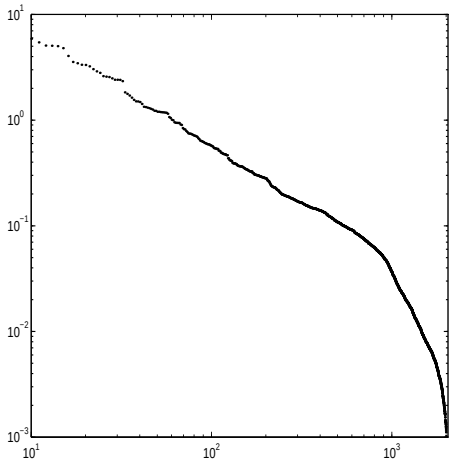

Figure 7.8: The wavelet coefficients of the function in Figure 7.7, in decreasing order of magnitude.

7.3 Problems Caused by the Choice of Spaces

What Does Unconditionality Cost?

We shall see in this section that images are sparse in Besov spaces. The norm equivalences between Besov norms and coefficients of wavelet expansions state that the unit ball B of a Besov space (in the equivalent norm in terms of the wavelet coefficients) is *orthosymmetric* [35] with respect to the axes given by the wavelet basis; that is, if $\sum c_{j,k}\psi_{j,k} \in B$, then $\sum \epsilon_{j,k} c_{j,k}\psi_{j,k} \in B$ for all $\epsilon_{j,k} \in \{-1, +1\}$. Perhaps more interestingly, there is more than just orthosymmetry: for any collection (σ_j) of permutations, $\sum c_{j,\sigma_j(k)}\psi_{j,k} \in B$ as well. That is, if we shuffle the positions of the coefficients randomly within every detail space, and build a new function out of that, this function is also in B.

It is natural to test to what extent the "class of images" shares these two properties. In Figure 7.9, we have the original Window image. We perform two operations on this image. The first is to change the sign of every coefficient in a random way. The resulting function for an outcome of this random change is shown in Figure 7.10. Note that the sharp edges are completely destroyed. The pointwise Hölder smoothness of each point should remain the same,[2] since it only depends on the sizes of coefficients. However, an edge (or even a jump discontinuity in one-dimension) is a particular C^0 singularity and is not invariant under a generic sign flip operation. Indeed, later in this section we prove a theorem on the asymptotic distribution of signs of wavelet coefficients that are in the cone of influence of a point at which there is a jump discontinuity, confirming this non-invariance.

The second operation involves a random shuffling of the wavelet coefficients within

[2]This is an approximate statement, see [36] for a more accurate characterization of local regularity in wavelet bases.

Figure 7.9: The Window image.

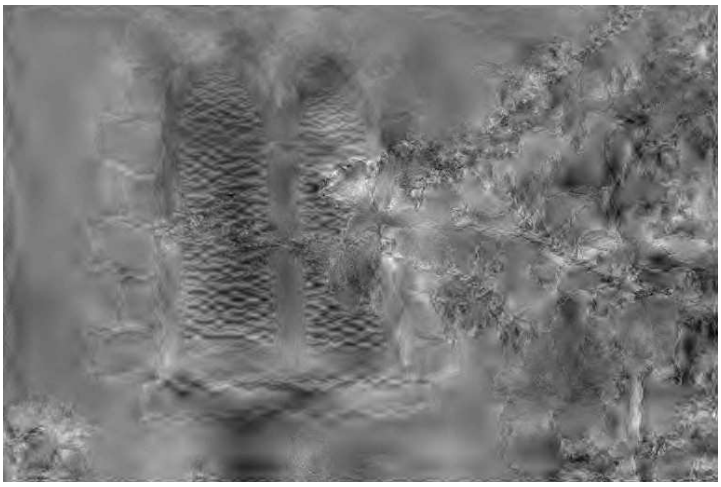

Figure 7.10: Window image after a random sign change.

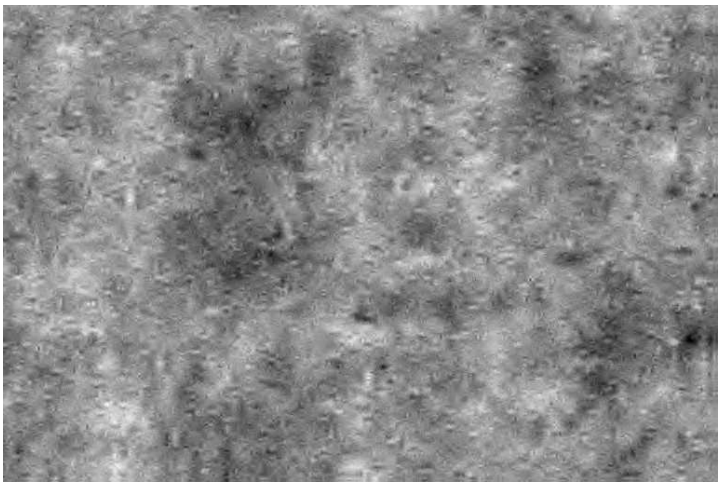

Figure 7.11: Window image after a random shuffling within every band.

each detail space. The result is shown in Figure 7.11. Note that this time the resulting image is not recognizable as a natural image (with the exception of modern art!). It is however interesting to note that this new image has the same equivalent Besov norm.

As a result of this experiment, we conclude that the class of images is "sparse" in any Besov space ball. By sparsity, we understand the lack of orthosymmetry and invariance of certain axes. Thus, a classification of images using only sizes of coefficients is apparently incomplete.

An Asymptotic Distribution Theorem for the Signs of Wavelet Coefficients

Let $f : \mathbb{R} \to \mathbb{R}$ be a C^α function for some $\alpha > 0$, except at the point x_0, where it has a jump singularity. Without loss of generality, we assume that $\alpha \leq 1$. Consider a wavelet basis $(\psi_{j,k})$ generated from a compactly supported mother wavelet ψ whose support is $[0, m]$ for some integer m. We also make the technical and not strict assumption that on its support, ψ vanishes only on a set of zero measure. Let the wavelet expansion of f be $\sum c_{j,k}\psi_{j,k}$ and denote by $\Lambda(x_0)$ the indices of wavelets that are in the *cone of influence* of x_0, defined by

$$\Lambda(x_0) = \{(j,k) : x_0 \in \text{supp}(\psi_{j,k})\}. \tag{7.8}$$

We will prove the following "structure theorem" regarding the asymptotic distribution of the signs of the wavelet coefficients $(c_\lambda)_{\lambda\in\Lambda(x_0)}$:

Theorem 7.1. *For almost every x_0, the signs of the wavelet coefficients $(c_\lambda)_{\lambda\in\Lambda(x_0)}$ of a function f that is C^α away from x_0 have a unique asymptotic distribution that depends only on the mother wavelet ψ and on $\sigma(x_0) := sgn(f(x_0-) - f(x_0+))$, i.e., the sign of the jump at x_0. More precisely, for each $\sigma \in \{-1, +1\}$, there exist p_1,*

$p_{-1} \in [0,1]$, *depending only on* ψ *and on* $\sigma(x_0)$, *such that* $p_{-1} + p_1 = 1$, *and*

$$\frac{\#\{(j,k) \in \Lambda(x_0) : j \le J, \;\; sgn(c_{j,k}) = \sigma\}}{\#\{(j,k) \in \Lambda(x_0) : j \le J\}} \longrightarrow p_\sigma. \tag{7.9}$$

In other words, the frequency of the occurence of the sign σ in the first J levels of the cone of influence of x_0 converges to an asymptotic value p_σ as $J \to \infty$. We use the convention $\mathrm{sgn}(0) = 1$.

Proof. Without loss of generality, and to simplify some of the notation, we first assume that $\alpha \le 1$. If $\alpha > 1$, then one simply replaces α by 1 in the analysis below.

We can safely assume that x_0 is not a dyadic rational number, so that $2^j x_0$ is never an integer. Then, $\Lambda(x_0)$ has the simple characterization

$$\begin{aligned} (j,k) \in \Lambda(x_0) \quad &\iff \quad 0 < 2^j x_0 - k < m \\ &\iff \quad k \in \{\lfloor 2^j x_0 \rfloor, \ldots, \lfloor 2^j x_0 \rfloor - (m-1)\}. \end{aligned} \tag{7.10}$$

This motivates us to define $k_{j,l} = \lfloor 2^j x_0 \rfloor - l$, for $l = 0, \ldots, m-1$. Let $[j,l]$ be a shorthand notation for $(j, k_{j,l})$. Thus,

$$\Lambda(x_0) = \{[j,l] : 1 \le j \le \infty, \;\; 0 \le l \le m-1\}. \tag{7.11}$$

For convenience, we write f as $f_1 \chi_{(-\infty,x_0]} + f_2 \chi_{(x_0,\infty)}$, where f_1 and f_2 are C^α everywhere, and where the choice for the value of f at x_0 is arbitrary. Let $\sum \tilde{c}_{j,k} \psi_{j,k}$ be the wavelet expansion of the piecewise constant function $\tilde{f}_{x_0} = f_1(x_0)\chi_{(-\infty,x_0]} + f_2(x_0)\chi_{(x_0,\infty)}$. We first calculate $\tilde{c}_{j,k}$ explicitly as follows:

$$\begin{aligned} \tilde{c}_{j,k} &= \langle f_1(x_0)\chi_{(-\infty,x_0]} + f_2(x_0)\chi_{(x_0,\infty)}, \psi_{j,k} \rangle \\ &= f_1(x_0) \int_{-\infty}^{x_0} \psi_{j,k}\, dx + f_2(x_0) \int_{x_0}^{\infty} \psi_{j,k}\, dx \\ &= [f_1(x_0) - f_2(x_0)] \Psi_{j,k}(x_0), \end{aligned} \tag{7.12}$$

where $\Psi_{j,k}$ is defined to be the primitive of $\psi_{j,k}$, and we have made use of the fact that $\int \psi\, dx = 0$. It is clear that $\Psi_{j,k} = 2^{-j/2}\Psi(2^j \cdot - k)$, where $\Psi(x) := \int_{-\infty}^{x} \psi(y)\, dy$. Note that $\mathrm{supp}(\Psi) = \mathrm{supp}(\psi) = [0,m]$. Now, by direct substitution,

$$\tilde{c}_{[j,l]} = 2^{-j/2}[f_1(x_0) - f_2(x_0)]\Psi(\langle 2^j x_0 \rangle + l), \tag{7.13}$$

where $\langle u \rangle := u \pmod 1 = u - \lfloor u \rfloor$ is a notation for the fractional part. We shall first prove that for each l, the sequence $(\mathrm{sgn}(\tilde{c}_{[j,l]}))_{j=1}^\infty$ has an asymptotic distribution. To prove this, we shall use the following lemma of Weyl, whose origin goes back to Hardy and Littlewood.

Lemma 7.2 (Weyl). $(\langle 2^j u \rangle)_{j=1}^\infty$ *is uniformly distributed for almost every* u.[3]

[3] Actually, a stronger result holds, due to Weyl [13, pp. 32]: For every *distinct* sequence of integers a_j, the sequence $(\langle a_j u \rangle)_{j=1}^\infty$ is uniformly distributed for almost every u.

For a proof of this lemma, see [13], page 32.

Now, let x_0 be such that $(\langle 2^j x_0\rangle)_{j=1}^{\infty}$ is uniformly distributed. Note that $\sigma(x_0) = \operatorname{sgn}(f_1(x_0) - f_2(x_0))$. Then, since $\operatorname{sgn}(\tilde{c}_{[j,l]}) = \sigma(x_0)\cdot\operatorname{sgn}(\Psi(\langle 2^j x_0\rangle + l)$, it follows that

$$\frac{1}{J}\#\{1 \le j \le J : \operatorname{sgn}(\tilde{c}_{[j,l]}) = \sigma\} \longrightarrow \operatorname{mes}\{\operatorname{sgn}(\Psi\cdot\chi_{[l,l+1]}) = \sigma\cdot\sigma(x_0)\} \tag{7.14}$$

as $J\to\infty$, where $\operatorname{mes}(A)$ denotes the Lebesgue measure of a (measurable) set A. It immediately follows by summing through $l = 0,\ldots,m-1$ that for $(\tilde{c}_{j,k})$, the quantity on the left hand side of (7.9) exists and is equal to

$$\frac{1}{m}\operatorname{mes}\{\operatorname{sgn}(\Psi) = \sigma\cdot\sigma(x_0)\}. \tag{7.15}$$

We conclude the proof by showing that the same asymptotic distribution holds for the signs of the actual coefficients $(c_{j,k})$. We shall do this by estimating the frequency of the event that $c_{[j,l]}$ and $\tilde{c}_{[j,l]}$ have opposite signs. Let

$$n_{J,l} := \frac{1}{J}\#\{1 \le j \le J : \operatorname{sgn}(c_{[j,l]}) \ne \operatorname{sgn}(\tilde{c}_{[j,l]})\}. \tag{7.16}$$

It is clear that $\operatorname{sgn}(c_{[j,l]}) \ne \operatorname{sgn}(\tilde{c}_{[j,l]})$ implies $|c_{[j,l]} - \tilde{c}_{[j,l]}| \ge |\tilde{c}_{[j,l]}|$. On the other hand, $|c_{[j,l]} - \tilde{c}_{[j,l]}|$ can be estimated by

$$\begin{aligned}
|c_{[j,l]} - \tilde{c}_{[j,l]}| &= |\langle f - \tilde{f}_{x_0}, \psi_{[j,l]}\rangle| \\
&\le \|(f - \tilde{f}_{x_0})\cdot\chi_{\operatorname{supp}(\psi_{[j,l]})}\|_{L^\infty}\|\psi_{[j,l]}\|_{L^1} \\
&\le \max(|f_1|_{C^\alpha}, |f_2|_{C^\alpha})\cdot(m2^{-j})^\alpha\cdot(m2^{-j})^{1/2} \\
&\le C\,2^{-j(\alpha+1/2)},
\end{aligned} \tag{7.17}$$

for a constant C that does not depend on j or l. Fix a positive integer J_0 and consider $J \ge J_0$. Then, by (7.12),

$$\begin{aligned}
n_{J,l} &\le \frac{1}{J}\#\{1 \le j \le J : 2^{-j/2}|f_1(x_0) - f_2(x_0)|\cdot|\Psi(\langle 2^j x_0\rangle + l)| \le C\,2^{-j(\alpha+1/2)}\} \\
&= \frac{1}{J}\#\{1 \le j \le J : |\Psi(\langle 2^j x_0\rangle + l)| \le C'\,2^{-j\alpha}\} \\
&\le \frac{1}{J}\left(J_0 + \#\{J_0 < j \le J : |\Psi(\langle 2^j x_0\rangle + l)| \le C'\,2^{-J_0\alpha}\}\right).
\end{aligned} \tag{7.18}$$

Here, $C' = C/|f_1(x_0) - f_2(x_0)|$. Finally, by first taking the lim sup of both sides as $J\to\infty$ and then the infimum over all J_0,

$$\begin{aligned}
\limsup_{J\to\infty} n_{J,l} &\le \inf_{J_0} \operatorname{mes}\{x \in [l, l+1] : |\Psi(x)| \le C'\,2^{-J_0\alpha}\} \\
&= \operatorname{mes}\{x \in [l, l+1] : \Psi(x) = 0\} \\
&= 0,
\end{aligned} \tag{7.19}$$

due to our assumption that ψ (and *a fortiori* Ψ) vanishes on $[0, m]$ only on a set of zero measure. This indeed implies that

$$\lim_{J\to\infty}\sum_{l=0}^{m-1} n_{J,l} = 0, \tag{7.20}$$

concluding the proof. □

Generalization to higher dimensions

We can generalize Theorem 7.1 to arbitrary dimensions. The setting is the following: f is a C^α function on $\mathbb{R}^d$ except on the surface M, which is a $C^{(1)}$ manifold of codimension 1. We assume that for each point $\mathbf{x}_0$ on M, the two limits of f at $\mathbf{x}_0$ obtained as one approaches in the direction of the unit normal $\mathbf{n}_0$ at $\mathbf{x}_0$ and in the opposite direction are distinct. Let these numbers be $f_1(\mathbf{x}_0)$ and $f_2(\mathbf{x}_0)$. Similar to the analysis above, we define an auxiliary function $\tilde{f}_{\mathbf{x}_0}$ by

$$\tilde{f}_{\mathbf{x}_0} = f_1(\mathbf{x}_0)\chi_{\tilde{\Omega}_1} + f_2(\mathbf{x}_0)\chi_{\tilde{\Omega}_2}, \tag{7.21}$$

where we set $\tilde{\Omega}_1 = \{\mathbf{x} : \mathbf{n}_0 \cdot (\mathbf{x} - \mathbf{x}_0) \geq 0\}$, and $\tilde{\Omega}_2 = \tilde{\Omega}_1^c$.

Let ψ be one of the $2^d - 1$ tensor product wavelets constructed from a univariate scaling function and the associated wavelet supported on $[0, m]$. For each j, the cone of influence $\Lambda(\mathbf{x}_0)$ at $\mathbf{x}_0$ will now contain a d-dimensional cubic array of indices with m^d elements. More precisely,

$$(j, \mathbf{k}) \in \Lambda(\mathbf{x}_0) \iff \mathbf{k} = \mathbf{k}_{j,\mathbf{l}} = \lfloor 2^j \mathbf{x_0} \rfloor - \mathbf{l} \quad \text{for some } \mathbf{l} \in \{0, \ldots, m-1\}^d, \tag{7.22}$$

where $\lfloor \cdot \rfloor$ is defined coordinatewise on a vector. We use the same shorthand notation $[j, \mathbf{l}]$ to denote $(j, \mathbf{k}_{j,\mathbf{l}})$.

Let $(c_{j,\mathbf{k}})$ and $(\tilde{c}_{j,\mathbf{k}})$ be the wavelet coefficients of f and $\tilde{f}_{\mathbf{x}_0}$ with respect to ψ. It follows by a straightforward calculation that

$$\tilde{c}_{[j,\mathbf{l}]} = 2^{-jd/2}[f_1(\mathbf{x}_0) - f_2(\mathbf{x}_0)]\, \Psi_{\mathbf{n}_0}(\langle 2^j \mathbf{x}_0 \rangle + \mathbf{l}), \tag{7.23}$$

where $\Psi_{\mathbf{n}_0}$ is defined by

$$\Psi_{\mathbf{n}_0}(\mathbf{s}) = \int_{(\mathbf{x}-\mathbf{s})\cdot\mathbf{n}_0 \geq 0} \psi(\mathbf{x}) d\mathbf{x}. \tag{7.24}$$

We define the "jump" at $\mathbf{x}_0$ to be $f_1(\mathbf{x}_0) - f_2(\mathbf{x}_0)$.

The rest of the proof consists of two ingredients. The first is the metric result stating that for almost all $\mathbf{u} \in \mathbb{R}^d$, the sequence $(\langle 2^j \mathbf{u} \rangle)_{j=1}^\infty$ is uniformly distributed mod 1 in $\mathbb{R}^d$ [13, pp. 52, Ex. 6.12]. As in the proof of Theorem 7.1, this immediately results in an asymptotic distribution property for $(\mathrm{sgn}(\tilde{c}_\lambda))_{\lambda \in \Lambda(\mathbf{x}_0)}$ that holds for almost every $\mathbf{x}_0$ and depends only on ψ, $\mathbf{n}_0$ and $\mathrm{sgn}(f_1(\mathbf{x}_0) - f_2(\mathbf{x}_0))$. The second ingredient is a size estimate for $|c_{[j,\mathbf{l}]} - \tilde{c}_{[j,\mathbf{l}]}|$. Let us see that

$$|c_{[j,\mathbf{l}]} - \tilde{c}_{[j,\mathbf{l}]}| \leq C\, 2^{-j(d/2+\alpha)}, \tag{7.25}$$

where we have again assumed that $\alpha \leq 1$. To prove this inequality, we first describe the manifold M around $\mathbf{x}_0$ by an associated $C^{(1)}$ function ϕ defined on a neighborhood U of $\mathbf{x}_0$: $M \cap U = \{\mathbf{x} \in U : \phi(\mathbf{x}) = 0\}$, and $\nabla\phi$ is nonvanishing on U. Then $\mathbf{n}_0 = \nabla\phi(\mathbf{x}_0)/|\nabla\phi(\mathbf{x}_0)|$ and the tangent plane at $\mathbf{x}_0$ is given by $T_p(\mathbf{x}_0) = \partial\tilde{\Omega}_1 = \{\mathbf{x} : \nabla\phi(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) = 0\}$. Consider a sufficiently small ball B_0 around $\mathbf{x}_0$ on which

$$\phi(\mathbf{x}) = \nabla\phi(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) + O(|\mathbf{x} - \mathbf{x}_0|^2) \tag{7.26}$$

(note that $\phi(\mathbf{x}_0) = 0$), and assume j is large enough to ensure that $I := \mathrm{supp}(\psi_{[j,\mathbf{l}]})$ is contained in B_0. Set $\Omega_1 := \{\mathbf{x} : \phi(\mathbf{x}) > 0\}$, $\Omega_2 := \{\mathbf{x} : \phi(\mathbf{x}) < 0\}$, and decompose $I \backslash M$ as

$$I \backslash M = (I \cap \Omega_1 \cap \tilde{\Omega}_1) \cup (I \cap \Omega_1 \cap \tilde{\Omega}_1^c) \cup (I \cap \Omega_2 \cap \tilde{\Omega}_2) \cup (I \cap \Omega_2 \cap \tilde{\Omega}_2^c). \tag{7.27}$$

Now, f is C^α on Ω_1 and Ω_2 so that

$$\begin{aligned} \int_{I \cap \Omega_1 \cap \tilde{\Omega}_1} |f - \tilde{f}|.|\psi_{[j,\mathbf{l}]}| d\mathbf{x} + \int_{I \cap \Omega_2 \cap \tilde{\Omega}_2} |f - \tilde{f}|.|\psi_{[j,\mathbf{l}]}| d\mathbf{x} &\leq 2\,\delta(I)^\alpha |f|_{C^\alpha} \|\psi_{[j,\mathbf{l}]}\|_{L^1} \\ &\leq C\, 2^{-j(d/2+\alpha)}, \end{aligned} \tag{7.28}$$

where $\delta(I) = 2^{-j} d^{1/2}$ is the diameter of I.

Next, we estimate the measures of the remaining sets. Let us do this for $I \cap \Omega_1 \cap \tilde{\Omega}_1^c$. If $\mathbf{x}$ is in $\Omega_1 \cap \tilde{\Omega}_1^c$, then $\phi(x) > 0$ but $\nabla\phi(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) < 0$. This, together with (7.26) implies that each of these numbers is bounded by $O(2^{-2j})$ in magnitude. Hence, $\mathrm{dist}(\mathbf{x}, T_p(\mathbf{x}_0)) = |\nabla\phi(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)| \leq C\, 2^{-2j}$. This yields to

$$\begin{aligned} \mathrm{mes}(I \cap \Omega_1 \cap \tilde{\Omega}_1^c) &\leq 2 \cdot \mathrm{Area}(T_p(\mathbf{x}_0) \cap I) \cdot \sup_{\mathbf{x} \in I \cap \Omega_1 \cap \tilde{\Omega}_1^c} \mathrm{dist}(\mathbf{x}, T_p(\mathbf{x}_0)) \\ &\leq C\, 2^{-j(d-1)} 2^{-2j}, \end{aligned} \tag{7.29}$$

which leads to the estimate

$$\begin{aligned} \int_{I \cap \Omega_1 \cap \tilde{\Omega}_1^c} |f - \tilde{f}|.|\psi_{[j,\mathbf{l}]}| d\mathbf{x} &\leq C\, 2^{-j(d+1)} \|f\|_{L^\infty} \|\psi_{[j,\mathbf{l}]}\|_{L^\infty} \\ &\leq C\, 2^{-j(d/2+1)}. \end{aligned} \tag{7.30}$$

One applies exactly the same argument to $I \cap \Omega_2 \cap \tilde{\Omega}_2^c$ and this finishes the proof of the estimate (7.25).

Following the same steps in the proof in the one dimensional case, these two ingredients would yield the same asymptotic distribution for $(\mathrm{sgn}(c_\lambda))_{\lambda \in \Lambda(\mathbf{x}_0)}$ provided that $\mathrm{mes}\{\mathbf{s} \in [0,m]^d : \Psi_{\mathbf{n}_0}(\mathbf{s}) = 0\} = 0$. We have thus proved:

Theorem 7.3. *For all $\mathbf{n}_0$ such that mes$\{\mathbf{s} \in [0,m]^d : \Psi_{\mathbf{n}_0}(\mathbf{s}) = 0\} = 0$, and for almost every $\mathbf{x}_0$, the signs of the wavelet coefficients $(c_\lambda)_{\lambda \in \Lambda(x_0)}$ of a function f that is C^α except on a $C^{(1)}$ surface $M \ni \mathbf{x}_0$ with the unit normal $\mathbf{n}_0$ at $\mathbf{x}_0$ have a unique asymptotic distribution that depends only on the mother wavelet ψ, $\mathbf{n}_0$ and the sign of the jump at $\mathbf{x}_0$.*

Remark: While it is true that $\Psi_{\mathbf{n}_0}$ may vanish identically for some $\mathbf{n}_0$, such as for some of the coordinate axis directions depending on which wavelet is considered among the possible $2^d - 1$, we conjecture that the set of directions $\mathbf{n}_0$ for which this is the case has measure zero.

7.4 A More Refined Approach: The Multifractal Formalism

We saw before that although global smoothness spaces are the fundamental spaces to study the (asymptotic) performance of approximation error, they may not provide a complete picture for classifying images in terms of their smoothness. A natural approach is to study the role of local regularity in this respect. In this section, we shall attempt to apply the multifractal formalism for functions to images. Our treatment will not always be fully rigorous and we shall often be satisfied with approximate quantities. We follow some of the notation and the definitions in [37].

We start with the definition of pointwise Hölder spaces. A function $f : \mathbb{R}^d \to \mathbb{R}$ belongs to the class $C^\alpha(x_0)$ if there exists a polynomial P_{x_0} whose degree is at most α, such that

$$|f(x) - P_{x_0}(x)| \leq C|x - x_0|^\alpha \tag{7.31}$$

on a neighborhood of x_0. P_{x_0} is usually the Taylor polynomial of f at x_0, although this may not always be the case. One defines the pointwise Hölder exponent $\alpha(x_0)$ of f at x_0 to be

$$\alpha(x_0) := \sup\{\alpha : f \in C^\alpha(x_0)\}. \tag{7.32}$$

A related space is $\Gamma^\alpha(x_0)$, which is defined to be

$$\Gamma^\alpha(x_0) = \bigcap_{\beta<\alpha} C^\beta(x_0) \setminus \bigcup_{\eta>\alpha} C^\eta(x_0), \tag{7.33}$$

i.e. $f \in \Gamma^\alpha(x_0)$ if and only if $\alpha(x_0) = \alpha$. Then the *singularity spectrum* (or the *Hölder spectrum*) of f is defined to be the function

$$D(\alpha) := \dim_{\text{Haus}}\{x_0 : f \in \Gamma^\alpha(x_0)\}, \tag{7.34}$$

i.e., the Hausdorff dimension of the set of points where the pointwise Hölder exponent is α. The Hausdorff dimension is sometimes replaced with the packing dimension or the box dimension (in computations). The following is the heuristics of the *structure function* method which is used to compute the singularity spectrum.

We start with the structure function defined as:

$$S_q(h) = \int |f(x+h) - f(x)|^q dx. \tag{7.35}$$

Then, $|f(x_0+h) - f(x_0)| \sim |h|^\alpha$ if $f \in \Gamma^\alpha(x_0)$. For each α, one approximates the set $\{x_0 : f \in \Gamma^\alpha(x_0)\}$ by a union of $|h|^{-D(\alpha)}$ cubes of size $|h|^d$, so that for small h,

$$S_q(h) \sim \sum_\alpha |h|^{q\alpha}|h|^{-D(\alpha)}|h|^d. \tag{7.36}$$

So, if $S_q(h)$ behaves as $|h|^{\zeta(q)}$ as $h \to 0$, then

$$\zeta(q) = \inf_\alpha (\alpha q + d - D(\alpha)), \tag{7.37}$$

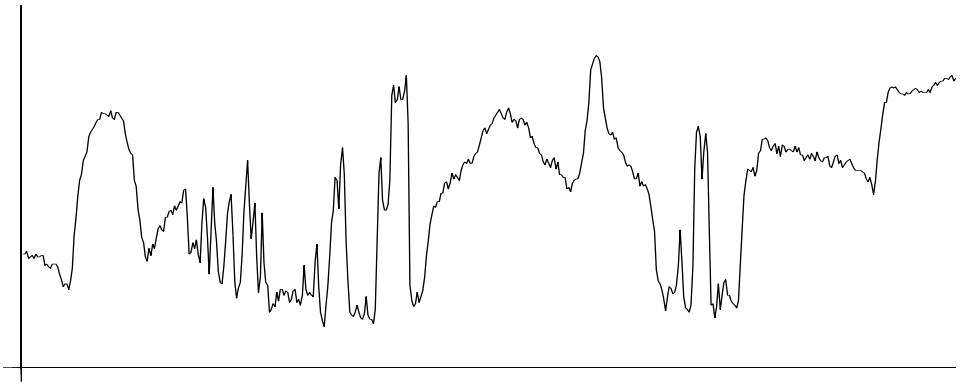

Figure 7.12: A sample line from the Lena image.

since this would be the dominant term in (7.36). One recognizes from (7.37) that ζ is the *Legendre transform* of D.

It is possible to invert the formula (7.37) when the function $D(\cdot)$ is concave. Then, the inverse Legendre transform has the same form:

$$D(\alpha) = \inf_q (\alpha q + d - \zeta(q)). \tag{7.38}$$

The multifractal formalism consists of the formulas (7.37) and (7.38). If $D(\cdot)$ is not concave, then the formula (7.38) recovers the concave majorant of $D(\cdot)$ only. Jaffard showed that under fairly general conditions, the singularity spectrum can be *any* function, so that (7.38) can easily fail. Thus, none of the two parts of the multifractal formalism has to hold in general. (See [37, 38] for an extensive treatment.)

Through the function $S_q(\cdot)$, it is possible to relate the singularity spectrum $D(\cdot)$ with the Besov spaces that contain f. Let $\alpha_f(p) := \sup\{\alpha : f \in B^\alpha_{p,\infty}\}$, as we defined in (7.1). If $S_q(h) \sim |h|^{\zeta(q)}$, then this means $\omega_1(f,h)_q \sim |h|^{\zeta(q)/q}$, so that $f \in B^{\zeta(q)/q}_{q,\infty}$. That is, $\alpha_f(q)$ is approximately equal to $\zeta(q)/q$.

To demonstrate this, we compute the singularity spectrum of a sample line taken from a natural image, shown in Figure 7.12. The corresponding singularity spectrum is plotted in Figure 7.13, and the "smoothness curve" in Figure 7.14.

Note that the structure function method, as given by (7.35), can work only for $\alpha < 1$. For the analysis of higher smoothness, one has to use higher order differences. An elegant way to do this, while at the same time introducing a stabilizing averaging process, is the *Wavelet Transform Integral* method, which computes

$$Z_q(a) = \int |(Wf)(a,b)|^q db, \tag{7.39}$$

where $(Wf)(a,b) = a^{-d} \int f(t)\psi(\frac{t-b}{a})dt$ is the continuous wavelet transform of f, for a wavelet ψ with sufficient smoothness and number of vanishing moments. The quantity $Z_q(a)$ is meant to replace $S_q(h)$: if $Z_q(a) \sim a^{\eta(q)}$, then $D(\alpha)$ is computed using the formula (7.38), with $\zeta(q)$ replaced by $\eta(q)$. For negative q, one clearly

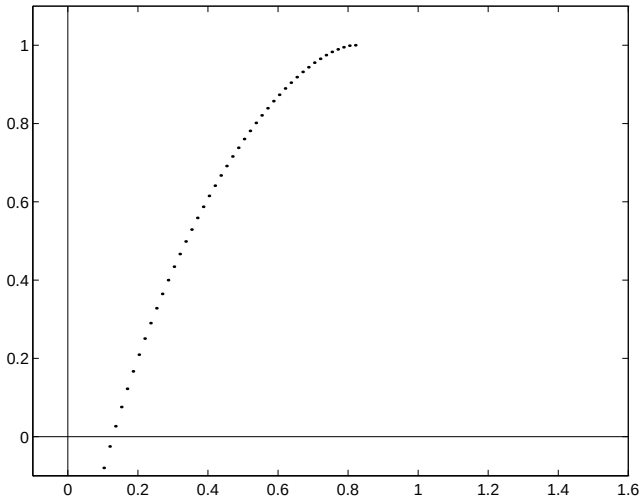

Figure 7.13: The singularity spectrum of the function in Figure 7.12, computed using the structure function method. The horizontal axis is α, and the vertical axis is $D(\alpha)$.

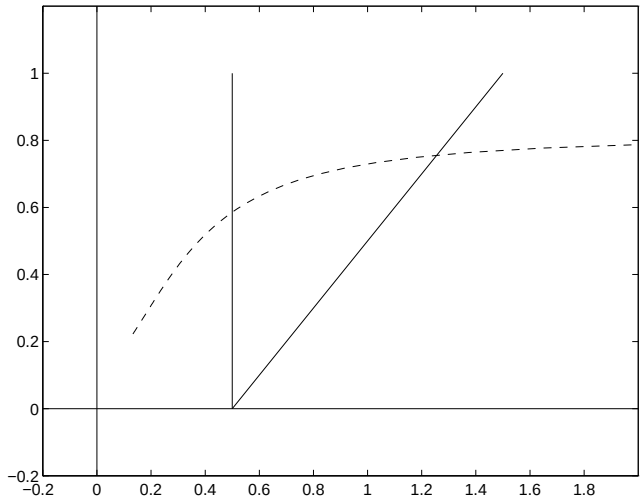

Figure 7.14: The smoothness curve for the function in Figure 7.12. The dotted curve plots $\alpha_f(q)$ as a function of the abscissa $1/q$.

has problems due to the zeros of the wavelet transform; a numerical ad-hoc method seeking to circumvent this is to use the *Wavelet Transform Modulus Maxima* method, which computes (7.39) not by integrating over the whole of the domain, but only as a summation over the lines of local maxima of the wavelet transform.

Note that although this multifractal formalism has been proposed with the goal of understanding local behavior, it still does so "on average" only, leading again to global estimates. The "equality" $\alpha_f(q) \approx \zeta(q)/q$ expresses this: even if we accept all the assumptions made by the multifractal formalism (which are very hard to verify in practice), then the structure function estimates, whether via (7.35) or the more refined (7.39), only give us an intersection of Besov spaces to which our signal or image belongs. It follows apart from all the natural reservations about unverifiable assumptions in this method, we still have not overcome the shortcomings of the Besov classes as a natural framework for images.

Chapter 8

Appendix: Multiresolution Approximation and Wavelets

The building block of a multiresolution analysis of $L^2(\mathbb{R})$ is a principal shift invariant space V_0, defined to be the closed L^2-span of the integer translates of a function φ. One also requires that this family $\{\varphi(\cdot - k)\}_{k\in\mathbb{Z}}$ constitutes a Riesz basis of V_0, i.e., $\|\sum_k c_k\varphi(\cdot - k)\|_{L^2} \asymp \|c\|_{l^2}$ for all $c = (c_k)_{k\in\mathbb{Z}} \in l^2$. Then, the associated multiresolution analysis is a sequence V_j of subspaces, all generated from V_0 by scaling:

$$V_j := \{f(2^j\cdot) \,:\, f \in V_0\}. \tag{8.1}$$

Hence φ is called the *scaling function*. A key requirement is that this family be nested, i.e., $V_j \subset V_{j+1}$ for all j, which reduces to $V_0 \subset V_1$ from the definition. In terms of φ, this means

$$\varphi(x) = \sum_k h_k\varphi(2x - k); \tag{8.2}$$

this equation is called the *refinement equation*. The function φ is then said to be *refinable*, and the sequence of coefficients (h_k) is called the *refinement mask*. The final requirement is that $L^2(\mathbb{R})$ is approximable from V_j, that is, for all f, $\mathrm{dist}(f, V_j) \to 0$ as $j \to \infty$.

If, for another refinable function $\tilde{\varphi}$, the family $\{\tilde{\varphi}(\cdot - k)\}_{k\in\mathbb{Z}}$ is biorthogonal to the family $\{\varphi(\cdot - k)\}_{k\in\mathbb{Z}}$, i.e., $\langle\varphi(\cdot - k), \tilde{\varphi}(\cdot - l)\rangle = \delta_{k,l}$ for all $k, l \in \mathbb{Z}$, then the operator $\mathbf{P}_{V_0} : L^2(\mathbb{R}) \to V_0$ defined by

$$\mathbf{P}_{V_0} f := \sum_k \langle f, \tilde{\varphi}(\cdot - k)\rangle\varphi(\cdot - k) \tag{8.3}$$

is a projection. $\tilde{\varphi}$ is called a *dual scaling function* for φ.

By scaling, $\{\varphi(2^j\cdot - k)\}_{k\in\mathbb{Z}}$ constitutes a Riesz basis for V_j. We normalize in L^2 and set $\varphi_{j,k}(\cdot) := 2^{j/2}\varphi(2^j\cdot - k)$, and define $\tilde{\varphi}_{j,k}$ similarly. The corresponding projection operator is defined by $\mathbf{P}_{V_j} f := \sum_k \langle f, \tilde{\varphi}_{j,k}\rangle\varphi_{j,k}$. It is natural to express $\mathbf{P}_{V_j} f$ as

$$\mathbf{P}_{V_j} f = \mathbf{P}_{V_0} f + \sum_{l=0}^{j-1} (\mathbf{P}_{V_{l+1}} f - \mathbf{P}_{V_l} f) \tag{8.4}$$

which results in the definition of the detail space $W_j := Ran\,(\mathbf{P}_{V_{j+1}} - \mathbf{P}_{V_j})$. Similarly, W_j is a scaled version of W_0. An important result is that W_0 is also a principal shift invariant space, spanned by $\{\psi(\cdot - k)\}_{k\in\mathbb{Z}}$ for a function $\psi \in V_1$, called the *mother wavelet*. The inclusion $W_0 \subset V_1$ yields

$$\psi(x) = \sum_k g_k \varphi(2x - k), \tag{8.5}$$

for some sequence of coefficients $g = (g_k)$. Similarly, there exists a dual wavelet $\tilde{\psi}$ such that the projector $\mathbf{P}_{W_j} := \mathbf{P}_{V_{j+1}} - \mathbf{P}_{V_j}$ can be expressed as

$$\mathbf{P}_{W_j} f = \sum_k \langle f, \tilde{\psi}_{j,k}\rangle \psi_{j,k}, \tag{8.6}$$

where $\tilde{\psi}_{j,k}$ and $\psi_{j,k}$ are defined analogously. The decomposition $f = \mathbf{P}_{V_0} f + \sum_{j=0}^{\infty} \mathbf{P}_{W_j} f$ leads to an expansion of f in wavelets:

$$f = \sum_k \langle f, \tilde{\varphi}_{0,k}\rangle \varphi_{0,k} + \sum_{j=0}^{\infty}\sum_k \langle f, \tilde{\psi}_{j,k}\rangle \psi_{j,k}. \tag{8.7}$$

If a scaling function is dual to itself, it is called orthogonal. In this case, $\mathbf{P}_{V_0}$ is an orthogonal projection and the mother wavelet is also dual to itself. On the other hand, there is always a particular choice for $\tilde{\varphi}$ that makes $\mathbf{P}_{V_0}$ an orthogonal projection. We do not necessarily require $\mathbf{P}_{V_0}$ to be orthogonal, but restrict to compactly supported scaling functions and wavelets [39, 40]. In this case, there are only finitely many nonzero coefficients in the refinement mask.

In the multidimensional setting, the multiresolution analysis of $L^2(\mathbb{R}^d)$ is constructed using the *tensor product strategy* in the following sense: Denote the scaling function φ by ψ^0, and the mother wavelet ψ by ψ^1. For each $\mathbf{i} = (i_1, \ldots, i_d) \in \{0,1\}^d$, define

$$\psi^{\mathbf{i}}(x_1, \ldots, x_d) := \psi^{i_1}(x_1) \cdots \psi^{i_d}(x_d). \tag{8.8}$$

Then $\psi^{\mathbf{0}}$ is the multivariate version of the scaling function φ, and the remaining $2^d - 1$ functions $\{\psi^{\mathbf{i}} : \mathbf{i} \in \{0,1\}^d \backslash \{\mathbf{0}\}\}$ are the wavelet functions in $\mathbb{R}^d$. We sometimes use the compact notations $\{\varphi_\lambda\}_{\lambda\in\Gamma_j}$, and $\{\psi_\lambda\}_{\lambda\in\Lambda_j}$ to denote the collections $\{\psi^{\mathbf{0}}(2^j \cdot - \mathbf{k}) : \mathbf{k} \in \mathbb{Z}^d\}$ and $\{\psi^{\mathbf{i}}(2^j \cdot - \mathbf{k}) : \mathbf{k} \in \mathbb{Z}^d,\ \mathbf{i} \in \{0,1\}^d \backslash \{\mathbf{0}\}\}$ of translated scaling functions and wavelets at scale 2^j (see [29]).

For an exquisite treatment of wavelet analysis and wavelet constructions, we refer to the book [41].

Bibliography

[1] S.R. Norsworthy, R. Schreier, and G.C. Temes, editors. *Delta-Sigma Data Converters*. IEEE Press, 1997.

[2] J.C. Candy and G.C. Temes, editors. *Oversampling Delta-Sigma Data Converters*. IEEE Press, 1992.

[3] Ingrid Daubechies and Ronald A. DeVore. Reconstructing a bandlimited function from very coarsely quantized data: I. a family of stable sigma-delta modulators of arbitrary order. *in preparation*, 1998.

[4] C.S. Güntürk, J. Lagarias, and V. Vaishampayan. On the robustness of single loop sigma-delta modulation. *to appear in IEEE Transactions on Information Theory.*, 2000.

[5] H. Helson. *Harmonic Analysis*. Addison-Wesley, 1983.

[6] H. Helson. On a theorem of Szegö. *Proc. Amer. Math. Soc*, 6:235–242, 1955.

[7] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford Univ. Press, 1962.

[8] Yitzhak Katznelson. *An Introduction to Harmonic Analysis*. Dover, 1976.

[9] Y. Meyer. *Wavelets and Operators*. Cambridge University Press, 1992.

[10] Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*. Springer-Verlag, 1993.

[11] Charles F. Osgood, editor. *Diophantine Approximation and its Applications*. Academic Press, 1973.

[12] M. Laczkovich. Discrepancy estimates for sets with small boundary. *Stud. Sci. Math. Hung.*, 30:105–109, 1995.

[13] L. Kuipers and H. Niederreiter. *Uniform Distribution of Sequences*. Wiley, 1974.

[14] M. Drmota and R.F. Tichy. *Sequences, Discrepancies and Applications*. Springer, 1997.

[15] H.L. Montgomery. *Ten Lectures on the Interface Between Analytic Number Theory and Harmonic Analysis*. AMS, 1994.

[16] E. Stein. *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*. Princeton University Press, 1993.

[17] S. Lang. *Introduction to Diophantine Approximation*. Springer-Verlag, 1995.

[18] R.M. Gray. Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input. *IEEE Transactions on Communications*, 37:588–599, June 1989.

[19] Nguyen T. Thao. personal communication. 2000.

[20] A. N. Kolmogorov and V. M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspehi*, 14(86):3–86, 1959 (also: AMS Translations, Series 2, Vol. 17, 1961, pp. 277-364).

[21] G. G. Lorentz, M. von Golitschek, and Y. Makovoz. *Constructive Approximation: Advanced Problems*. Springer-Verlag, 1993.

[22] S. Konyagin. personal communication. 1999.

[23] R. Calderbank and I. Daubechies. Shortcomings of democracy. *preprint*, 2000.

[24] S. Hein and A. Zakhor. Optimal decoding for data acquision applications of sigma delta modulators. *IEEE Transactions on Signal Processing*, 41:602–616, Feb 1993.

[25] I. Daubechies, R. DeVore, S. Güntürk, and V. Vaishampayan. Quantization error correction in a/d conversion. *in preparation*, 2000.

[26] O. Feely and L. O. Chua. The effect of integrator leak in $\Sigma - \Delta$ modulation. *IEEE Transactions on Circuits and Systems*, 38:1293–1305, Nov 1991.

[27] Ronald A. DeVore, Björn Jawerth, and Bradley J. Lucier. Image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, 38(2):719–746, March 1992.

[28] Ronald A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[29] A. Cohen. Wavelet methods in numerical analysis. In P.G.Ciarlet and J.L.Lions, editors, *Handbook of Numerical Analysis*, volume 7. Elsevier Science, 1999.

[30] Vladimir N. Temlyakov. The best m-term approximation and greedy algorithms. *Adv. Comput. Math.*, 3:249–265, 1998.

[31] Emmanuel J. Candès and David L. Donoho. Curvelets - a surprisingly effective nonadaptive representation for objects with edges. *preprint*, 1999.

[32] A. Cohen, Ronald A. DeVore, P. Petrushev, and H.Xu. Nonlinear approximation and the space BV($\mathbb{R}^2$). *Amer. J. Math.*, 121:587–628, 1999.

[33] Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

[34] Albert Cohen and Jean-Pierre d'Ales. Nonlinear approximation of random functions. *SIAM Journal of Applied Mathematics*, 57(2):518–540, April 1997.

[35] David L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. and Comp. Harm. Anal.*, 1:110–115, 1993.

[36] S. Jaffard and Y. Meyer. *Wavelet Methods for Pointwise Regularity and Local Oscillation of Functions*. AMS, 1996.

[37] S. Jaffard. Multifractal formalism for functions part I: Results valid for all functions. *SIAM J. Math. Anal.*, 28:944–970, 1997.

[38] S. Jaffard. Multifractal formalism for functions part II: Self-similar functions. *SIAM J. Math. Anal.*, 28:971–998, 1997.

[39] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math*, 41:909–996, 1988.

[40] A. Cohen, I. Daubechies, and J. C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math*, 45:485–560, 1992.

[41] Ingrid Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.