Lecture Notes and Background Materials for Math 5467: Introduction to the Mathematics of Wavelets

Willard Miller

May 3, 2006

Contents

List of Figures

Comment These are lecture notes for the course, and also contain background material that I won't have time to cover in class. I have included this supplementary material, for those students who wish to delve deeper into some of the topics mentioned in class.

Chapter 1

Introduction (from a signal processing point of view)

Let $f(t)$ be a real-valued function defined on the real line R and square integrable:

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

Think of $f(t)$ as the value of a signal at time t. We want to analyze this signal in ways other than the time-value form $t \rightarrow f(t)$ given to us. In particular we will analyze the signal in terms of frequency components and various combinations of time and frequency components. Once we have analyzed the signal we may want to alter some of the component parts to eliminate some undesirable features or to compress the signal for more efficient transmission and storage. Finally, we will reconstitute the signal from its component parts.

The three steps are:

- **Analysis.** Decompose the signal into basic components. We will think of the signal space as a vector space and break it up into a sum of subspaces, each of which captures a special feature of a signal.

- **Processing** Modify some of the basic components of the signal that were obtained through the analysis. Examples:

  1. audio compression
  2. video compression
  3. denoising

4. edge detection

- **Synthesis** Reconstitute the signal from its (altered) component parts. An important requirement we will make is **perfect reconstruction**. If we don't alter the component parts, we want the synthesized signal to agree exactly with the original signal. We will also be interested in the convergence properties of an altered signal with respect to the original signal, e.g., how well a reconstituted signal, from which some information may have been dropped, approximates the original signal.

Remarks:

- Some signals are discrete, e.g., only given at times $t_j = j, \quad j = 0, \pm 1, \pm 2, \cdots$. We will represent these as step functions.

- Audio signals (telephone conversations) are of arbitrary length but video signals are of fixed finite length, say 2π. Thus a video signal can be represented by a function $f(t)$ defined for $-\pi \le t < \pi$. Mathematically, we can extend f to the real line by requiring that it be periodic

$$f(t) = f(t + 2\pi)$$

  or that it vanish outside the interval $-\pi \le t < \pi$.

We will look at several methods for signal analysis:

- Fourier series
- The Fourier integral
- Windowed Fourier transforms (briefly)
- Continuous wavelet transforms (briefly)
- Filter banks
- Discrete wavelet transforms (Haar and Daubechies wavelets)

Mathematically, all of these methods are based on the decomposition of the Hilbert space of square integrable functions into orthogonal subspaces. We will first review a few ideas from the theory of vector spaces.

Chapter 2

Vector Spaces with Inner Product.

2.1 Definitions

Let F be either the field of real numbers R or the field of complex number C.

Definition 1 *A vector space V over F is a collection of elements (vectors) with the following properties:*

- *For every pair $u, v \in V$ there is defined a unique vector $w = u + v \in V$ (the sum of u and v)*

- *For every $\alpha \in F$, $u \in V$ there is defined a unique vector $z = \alpha u \in V$ (product of α and u)*

- *Commutative, Associative and Distributive laws*

  1. $u + v = v + u$
  2. $(u + v) + w = u + (v + w)$
  3. *There exists a vector $\Theta \in V$ such that $u + \Theta = u$ for all $u \in V$*
  4. *For every $u \in V$ there is a $-u \in V$ such that $u + (-u) = \Theta$*
  5. $1u = u$ *for all* $u \in V$
  6. $\alpha(\beta u) = (\alpha\beta)u$ *for all* $\alpha, \beta \in F$
  7. $(\alpha + \beta)u = \alpha u + \beta u$
  8. $\alpha(u + v) = \alpha u + \alpha v$

Definition 2 *A non-empty set W in V is a subspace of V if $\alpha u + \beta v \in W$ for all $\alpha, \beta \in F$ and $u, v \in W$.*

Note that W is itself a vector space over F.

Lemma 1 *Let $u_1, u_2, \cdots, u_m$ be a set of vectors in the vector space V. Denote by $[u_1, u_2, \cdots, u_m]$ the set of all vectors of the form $\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_m u_m$ for $\alpha_i \in F$. The set $[u_1, u_2, \cdots, u_m]$ is a subspace of V.*

PROOF: Let $u, v \in [u_1, u_2, \cdots, u_m]$. Thus,

$$u = \sum_{i=1}^{m} \alpha_i u_i, \qquad v = \sum_{i=1}^{m} \beta_i u_i$$

so

$$\alpha u + \beta v = \sum_{i=1}^{m} (\alpha \alpha_i + \beta \beta_i) u_i \in [u_1, u_2, \cdots, u_m].$$

Q.E.D.

Definition 3 *The elements $u_1, u_2, \cdots, u_p$ of V are linearly independent if the relation $\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_p u_p = \Theta$ for $\alpha_i \in F$ holds only for $\alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$. Otherwise $u_1, \cdots, u_p$ are linearly dependent*

Definition 4 *V is n-dimensional if there exist n linearly independent vectors in V and any $n+1$ vector in V are linearly dependent.*

Definition 5 *V is finite-dimensional if V is n-dimensional for some integer n. Otherwise V is infinite dimensional.*

Remark: If there exist vectors $u_1, \cdots, u_n$, linearly independent in V and such that every vector $u \in V$ can be written in the form

$$u = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n, \qquad \alpha_i \in F,$$

($\{u_1, \cdots, u_n\}$ *spans* V), then V is n-dimensional. Such a set $\{u_1, \cdots, u_n\}$ is called a *basis* for V.

Theorem 1 *Let V be an n-dimensional vector space and $u_1, \cdots, u_n$ a linearly independent set in V. Then $u_1, \cdots, u_n$ is a basis for V and every $u \in V$ can be written uniquely in the form*

$$u = \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_n u_n.$$

PROOF: let $u \in V$. then the set $u_1, \cdots, u_n, u$ is linearly dependent. Thus there exist $\alpha_1, \cdots, \alpha_{n+1} \in F$, not all zero, such that

$$\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n + \alpha_{n+1} u = \Theta.$$

If $\alpha_{n+1} = 0$ then $\alpha_1 = \cdots = \alpha_n = 0$. Impossible! Therefore $\alpha_{n+1} \neq 0$ and

$$u = \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_n u_n, \qquad \beta_i = -\frac{\alpha_i}{\alpha_{n+1}}.$$

Now suppose

$$u = \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_n u_n = \gamma_1 u_1 + \gamma_2 u_2 + \cdots + \gamma_n u_n.$$

Then

$$(\beta_1 - \gamma_1) u_1 + \cdots + (\beta_n - \gamma_n) u_n = \Theta.$$

But the u_i form a linearly independent set, so $\beta_1 - \gamma_1 = 0, \cdots, \beta_n - \gamma_n = 0$. Q.E.D.

Examples 1 • *V_n, the space of all (real or complex) n-tuples $(\alpha_1, \cdots, \alpha_n)$, $\alpha_i \in F$. Here, $\Theta = (0, \cdots, 0)$. A standard basis is:*

$$u_1 = (1, 0 \cdots, 0), \quad u_2 = (0, 1, 0, \cdots, 0), \cdots, u_n = (0, 0, \cdots, 1).$$

PROOF:

$$(\alpha_1, \cdots, \alpha_n) = \alpha_1 u_1 + \cdots + \alpha_n u_n,$$

so the vectors span. They are linearly independent because

$$(\beta_1, \cdots, \beta_n) = \beta_1 u_1 + \cdots + \beta_n u_n = \Theta = (0, \cdots, 0)$$

if and only if $\beta_1 = \cdots = \beta_n = 0$. Q.E.D.

- *V_∞, the space of all (real or complex) infinity-tuples*

$$(\alpha_1, \alpha_2, \cdots, \alpha_n, \cdots).$$

This is an infinite-dimensional space.

- $C^{(n)}[a,b]$: *Set of all complex-valued functions with continuous derivatives of orders* $0,1,2,\cdots n$ *on the closed interval* $[a,b]$ *of the real line. Let* $t \in [a,b]$, *i.e.,* $a \leq t \leq b$ *with* $a < b$. *Vector addition and scalar multiplication of functions* $u, v \in C^{(n)}[a,b]$ *are defined by*

  $$[u+v](t) = u(t) + v(t) \qquad [\alpha u](t) = \alpha u(t).$$

  The zero vector is the function $\Theta(t) \equiv 0$. *The space is infinite-dimensional.*

- $S(J)$: *Space of all complex-valued step functions on the (bounded or unbounded) interval* J *on the real line.* s *is a step function on* J *if there are a finite number of non-intersecting bounded intervals* $J_1, \cdots, J_m$ *and complex numbers* $c_1, \cdots, c_m$ *such that* $s(t) = c_k$ *for* $t \in J_k$, $k = 1, \cdots, m$ *and* $s(t) = 0$ *for* $t \in J - \cup_{k=1}^{m} J_k$. *Vector addition and scalar multiplication of step functions* $s_1, s_2 \in S(J)$ *are defined by*

  $$[s_1 + s_2](t) = s_1(t) + s_2(t) \qquad [\alpha s_1](t) = \alpha s_1(t).$$

  (One needs to check that $s_1 + s_2$ *and* αs_1 *are step functions.) The zero vector is the function* $\Theta(t) \equiv 0$. *The space is infinite-dimensional.*

2.2 Schwarz inequality

Definition 6 *A vector space* $\mathcal{N}$ *over* F *is a normed linear space (pre-Banach space) if to every* $u, \in \mathcal{N}$ *there corresponds a real scalar* $||u||$ *such that*

1. $||u|| \geq 0$ *and* $||u|| = 0$ *if and only if* $u = 0$.
2. $||\alpha u|| = |\alpha| \; ||u||$ *for all* $\alpha \in F$.
3. *Triangle inequality.* $||u+v|| \leq ||u|| + ||v||$ *for all* $u, v \in \mathcal{N}$.

Examples 2

- $C^{(n)}[a,b]$: *Set of all complex-valued functions with continuous derivatives of orders* $0,1,2,\cdots n$ *on the closed interval* $[a,b]$ *of the real line. Let* $t \in [a,b]$, *i.e.,* $a \leq t \leq b$ *with* $a < b$. *Vector addition and scalar multiplication of functions* $u, v \in C^{(n)}[a,b]$ *are defined by*

  $$[u+v](t) = u(t) + v(t) \qquad [\alpha u](t) = \alpha u(t).$$

  The zero vector is the function $\Theta(t) \equiv 0$. *The norm is defined by* $||u|| = \int_a^b |u(t)| \, dt$.

- $S^1(J)$*: Set of all complex-valued step functions on the (bounded or unbounded) interval* J *on the real line.* s *is a step function on* J *if there are a finite number of non-intersecting bounded intervals* $J_1, \cdots, J_m$ *and real numbers* $c_1, \cdots, c_m$ *such that* $s(t) = c_k$ *for* $t \in J_k$*,* $k = 1, \cdots, m$ *and* $s(t) = 0$ *for* $t \in J - \cup_{k=1}^m J_k$*. Vector addition and scalar multiplication of step functions* $s_1, s_2 \in S(J)$ *are defined by*

  $$[s_1 + s_2](t) = s_1(t) + s_2(t) \qquad [\alpha s_1](t) = \alpha s_1(t).$$

  (One needs to check that $s_1 + s_2$ *and* αs_1 *are step functions.) The zero vector is the function* $\Theta(t) \equiv 0$*. The space is infinite-dimensional. We define the integral of a step function as the "area under the curve",i.e.,* $\int_J s(t)dt \equiv \sum_{k=1}^m c_k \ell(J_k)$ *where* $\ell(J_k)$ *= length of* J_k *=* $b - a$ *if* $J_k = [a, b]$ *or* $[a, b)$*, or* $(a, b]$ *or* (a, b)*. Note that*

  1. $s \in S(J) \Longrightarrow |s| \in S(J)$.
  2. $|\int_J s(t)dt| \leq \int_J |s(t)|dt$.
  3. $s_1, s_2 \in S(J) \Longrightarrow \alpha_1 s_1 + \alpha_2 s_2 \in S(J)$ *and* $\int_J (\alpha_1 s_1 + \alpha_2 s_2)(t)dt = \alpha_1 \int_J s_1(t)dt + \alpha_2 \int_J s2(t)dt$.

  Now we define the norm by $||s|| = \int_J |s(t)|dt$*. Finally, we adopt the rule that we identify* $s_1, s_2 \in S(J)$*,* $s_1 \sim s_2$ *if* $s_1(t) = s_2(t)$ *except at a finite number of points. (This is needed to satisfy property 1. of the norm.) Now we let* $S^1(J)$ *be the space of equivalence classes of step functions in* $S(J)$*. Then* $S^1(J)$ *is a normed linear space with norm* $|| \cdot ||$*.*

Definition 7 *A vector space* $\mathcal{H}$ *over* F *is an inner product space (pre-Hilbert space) if to every ordered pair* $u, v \in \mathcal{H}$ *there corresponds a scalar* $(u, v) \in F$ *such that*

Case 1*:* $F = C$*, Complex field*

- $(u, v) = \overline{(v, u)}$
- $(u + v, w) = (u, w) + (v, w)$
- $(\alpha u, v) = \alpha(u, v)$*, for all* $\alpha \in C$
- $(u, u) \geq 0$*, and* $(u, u) = 0$ *if and only if* $u = 0$

Note: $(u, \alpha v) = \bar{\alpha}(u, v)$

Case 2: $F = R$, *Real field*

1. $(u, v) = (v, u)$
2. $(u + v, w) = (u, w) + (v, w)$
3. $(\alpha u, v) = \alpha(u, v)$, *for all* $\alpha \in R$
4. $(u, u) \geq 0$, *and* $(u, u) = 0$ *if and only if* $u = 0$

Note: $(u, \alpha v) = \alpha(u, v)$

Unless stated otherwise, we will consider complex inner product spaces from now on. The real case is usually an obvious restriction.

Definition 8 *let $\mathcal{H}$ be an inner product space with inner product (u, v). The norm $||u||$ of $u \in \mathcal{H}$ is the non-negative number $||u|| = \sqrt{(u, u)}$.*

Theorem 2 *Schwarz inequality. Let $\mathcal{H}$ be an inner product space and $u, v \in \mathcal{H}$. Then*

$$|(u, v)| \leq ||u|| \; ||v||.$$

Equality holds if and only if u, v are linearly dependent.

PROOF: We can suppose $u, v \neq \Theta$. Set $w = u + \alpha v$, for $\alpha \in C$. The $(w, w) \geq 0$ and $= 0$ if and only if $u + \alpha v = 0$. hence

$$(w, w) = (u + \alpha v, u + \alpha v) = ||u||^2 + |\alpha|^2 \; ||v||^2 + \alpha(v, u) + \bar{\alpha}(u, v) \geq 0.$$

Set $\alpha = -(u, v)/||v||^2$. Then

$$||u||^2 + \frac{|(u, v)|^2}{||v||^2} - 2\frac{|(u, v)|^2}{||v||^2} \geq 0.$$

Thus $|(u, v)|^2 \leq ||u||^2 \; ||v||^2$. Q.E.D.

Theorem 3 *Properties of the norm. Let $\mathcal{H}$ be an inner product space with inner product (u, v). Then*

- $||u|| \geq 0$ *and* $||u|| = 0$ *if and only if* $u = 0$.

- $||\alpha u|| = |\alpha| \, ||u||$.
- *Triangle inequality.* $||u+v|| \leq ||u|| + ||v||$. *PROOF:*

$$||u+v||^2 = (u+v, u+v) = ||u||^2 + (u,v) + (v,u) + ||v||^2$$

$$\leq ||u||^2 + 2||u|| \, ||v|| + ||v||^2 = (||u|| + ||v||)^2.$$

Examples:

- $\mathcal{H}_n$ This is the space of complex n-tuples V_n with inner product

  $$(u,v) = \sum_{i=1}^{n} \alpha_i \overline{\beta}_i$$

  for vectors

  $$u = (\alpha_1, \cdots, \alpha_n), \qquad v = (\beta_1, \cdots, \beta_n), \qquad \alpha_i, \beta_i \in C.$$

- R_n This is the space of real n-tuples V_n with inner product

  $$(u,v) = \sum_{i=1}^{n} \alpha_i \beta_i$$

  for vectors

  $$u = (\alpha_1, \cdots, \alpha_n), \qquad v = (\beta_1, \cdots, \beta_n), \qquad \alpha_i, \beta_i \in R.$$

  Note that (u,v) is just the dot product. In particular for R_3 (Euclidean 3-space) $(u,v) = ||u|| \, ||v|| \cos\phi$ where $||u|| = \sqrt{\alpha_i^2 + \alpha_2^2 + \alpha_3^2}$ (the length of u), and $\cos\phi$ is the cosine of the angle between vectors u and v. The triangle inequality $||u+v|| \leq ||u|| + ||v||$ says in this case that the length of one side of a triangle is less than or equal to the sum of the lengths of the other two sides.

- $\hat{\mathcal{H}}_\infty$, the space of all complex infinity-tuples

  $$u = (\alpha_1, \alpha_2, \cdots, \alpha_n, \cdots).$$

  such that only a finite number of the α_i are nonzero. $(u,v) = \sum_{i=1}^{\infty} \alpha_i \overline{\beta}_i$.

- $\mathcal{H}_\infty$, the space of all complex infinity-tuples

  $$u = (\alpha_1, \alpha_2, \cdots, \alpha_n, \cdots).$$

  such that $\sum_{i=1}^{\infty} |\alpha_i|^2 < \infty$. Here, $(u, v) = \sum_{i=1}^{\infty} \alpha_i \overline{\beta}_i$. (need to verify that this is a vector space.)

- ℓ^2, the space of all complex infinity-tuples

  $$u = (\cdots, \alpha_{-1}, \alpha_0, \alpha_1, \cdots, \alpha_n, \cdots).$$

  such that $\sum_{i=-\infty}^{\infty} |\alpha_i|^2 < \infty$. Here, $(u, v) = \sum_{i=-\infty}^{\infty} \alpha_i \overline{\beta}_i$. (need to verify that this is a vector space.)

- $C_2^{(n)}[a, b]$: Set of all complex-valued functions $u(t)$ with continuous derivatives of orders $0, 1, 2, \cdots n$ on the closed interval $[a, b]$ of the real line. We define an inner product by

  $$(u, v) = \int_a^b u(t)\overline{v}(t)\, dt, \qquad u, v \in C_2^{(n)}[a, b].$$

- $C_2^{(n)}(a, b)$: Set of all complex-valued functions $u(t)$ with continuous derivatives of orders $0, 1, 2, \cdots n$ on the open interval (a, b) of the real line, such that $\int_a^b |u(t)|^2\, dt < \infty$, (Riemann integral). We define an inner product by

  $$(u, v) = \int_a^b u(t)\overline{v}(t)\, dt, \qquad u, v \in C_2^{(n)}(a, b).$$

  Note: $u(t) = t^{-1/3} \in C_2^{(2)}(0, 1)$, but $v(t) = t^{-1}$ doesn't belong to this space.

- $L_0^2[a, b]$: Set of all complex-valued functions $u(t)$ on the closed interval $[a, b]$ of the real line, such that $\int_a^b |u(t)|^2\, dt < \infty$, (Riemann integral). We define an inner product by

  $$(u, v) = \int_a^b u(t)\overline{v}(t)\, dt, \qquad u, v \in L^2[a, b].$$

  Note: There are problems here. Strictly speaking, this isn't an inner product. Indeed the nonzero function $u(0) = 1, u(t) = 0$ for $t > 0$ belongs to $L_0^2[0, 1]$, but $||u|| = 0$. However the other properties of the inner product hold.

- $S^2(J)$: Space of all complex-valued step functions on the (bounded or unbounded) interval J on the real line. s is a *step function* on J if there are a finite number of non-intersecting bounded intervals $J_1, \cdots, J_m$ and numbers $c_1, \cdots, c_m$ such that $s(t) = c_k$ for $t \in J_k$, $k = 1, \cdots, m$ and $s(t) = 0$ for $t \in J - \cup_{k=1}^m$. Vector addition and scalar multiplication of step functions $s_1, s_2 \in S(J)$ are defined by

  $$[s_1 + s_2](t) = s_1(t) + s_2(t) \quad [\alpha s_1](t) = \alpha s_1(t).$$

  (One needs to check that $s_1 + s_2$ and αs_1 are step functions.) The zero vector is the function $\Theta(t) \equiv 0$. Note also that the product of step functions, defined by $s_1 s_2(t) \equiv s_1(t)s_2(t)$ is a step function, as are $|s_1|$ and $\bar{s_1}$. We define the integral of a step function as $\int_J s(t)dt \equiv \sum_{k=1}^m c_k \ell(J_k)$ where $\ell(J_k)$ = length of J_k = $b - a$ if $J_k = [a, b]$ or $[a, b)$, or $(a, b]$ or (a, b). Now we define the inner product by $(s_1, s_2) = \int_J s_1(t)\overline{s_2(t)}dt$. Finally, we adopt the rule that we identify $s_1, s_2 \in S(J)$, $s_1 \sim s_2$ if $s_1(t) = s_2(t)$ except at a finite number of points. (This is needed to satisfy property 4. of the inner product.) Now we let $S^2(J)$ be the space of equivalence classes of step functions in $S(J)$. Then $S^2(J)$ is an inner product space.

2.3 An aside on completion of inner product spaces

This is supplementary material for the course. For motivation, consider the space R of the real numbers. You may remember from earlier courses that R can be constructed from the more basic space R' of rational numbers. The norm of a rational number r is just the absolute value $|r|$. Every rational number can be expressed as a ratio of integers $r = n/m$. The rationals are closed under addition, subtraction, multiplication and division by nonzero numbers. Why don't we stick with the rationals and not bother with real numbers? The basic problem is that we can't do analysis (calculus, etc.) with the rationals because they are not closed under limiting processes. For example $\sqrt{2}$ wouldn't exist. The Cauchy sequence $1, 1.4, 1.41, 1.414, \cdots$ wouldn't diverge, but would fail to converge to a rational number. There is a "hole" in the field of rational numbers and we label this hole by $\sqrt{2}$. We say that the Cauchy sequence above and all other sequences approaching the same hole are converging to $\sqrt{2}$. Each hole can be identified with the equivalence class of Cauchy sequences approaching the hole. The reals are just the space of equivalence classes of these sequences with appropriate definitions for addition and multiplication. Each rational number r corresponds to a constant

Cauchy sequence $r, r, r, \cdots$ so the rational numbers can be embedded as a subset of the reals. Then one can show that the reals are *closed*: every Cauchy sequence of real numbers converges to a real number. We have filled in all of the holes between the rationals. The reals are the *closure* of the rationals.

The same idea works for inner product spaces and it also underlies the relation between the Riemann integral of your calculus classes and the Lebesgue integral. To see how this goes, it is convenient to introduce the simple but general concept of a metric space. We will carry out the basic closure construction for metric spaces and then specialize to inner product and normed spaces.

Definition 9 *A set $\mathcal{M}$ is called a metric space if for each $u, v \in \mathcal{M}$ there is a real number $\rho(u, v)$ (the metric) such that*

1. $\rho(u, v) \geq 0, \qquad \rho(u, v) = 0$ *if and only if* $u = v$
2. $\rho(u, v) = \rho(v, u)$
3. $\rho(u, w) \leq \rho(u, v) + \rho(v, w)$ *(triangle inequality).*

REMARK: Normed spaces are metric spaces: $\rho(u, v) = ||u - v||$.

Definition 10 *A sequence $u_1, u_2, \cdots$ in $\mathcal{M}$ is called a Cauchy sequence if for every $\epsilon > 0$ there exists an integer $N(\epsilon)$ such that $\rho(u_n, u_m) < \epsilon$ whenever $n, m > N(\epsilon)$.*

Definition 11 *A sequence $u_1, u_2, \cdots$ in $\mathcal{M}$ is convergent if for every $\epsilon > 0$ there exists an integer $M(\epsilon)$ such that $\rho(u_n, u) < \epsilon$ whenever $n, m > M(\epsilon)$. here u is the limit of the sequence, and we write $u = \lim_{n\to\infty} u_n$.*

Lemma 2 *1) The limit of a convergent sequence is unique.*
2) Every convergent sequence is Cauchy.

PROOF: 1) Suppose $u = \lim_{n\to\infty} u_n$, $v = \lim_{n\to\infty} u_n$. Then $\rho(u, v) \leq \rho(u, u_n) + \rho(u_n, v) \to 0$ as $n \to \infty$. Therefore $\rho(u, v) = 0$, so $u = v$. 2) $\{u_n\}$ converges to u implies $\rho(u_n, u_m) \leq \rho(u_n, u) + \rho(u_m, u) \to 0$ as $n, m \to \infty$. Q.E.D

Definition 12 *A metric space $\mathcal{M}$ is complete if every Cauchy sequence in $\mathcal{M}$ converges.*

Examples 3 *Some examples of Metric spaces:*

- *Any normed space.* $\rho(u, v) = ||u - v||$. *Finite-dimensional inner product spaces are complete.*
- $\mathcal{M}$ *as the set of all rationals on the real line.* $\rho(u, v) = |u - v|$ *for rational numbers* u, v. *(absolute value) Here* $\mathcal{M}$ *is not complete.*

Definition 13 *A subset* $\mathcal{M}'$ *of the metric space* $\mathcal{M}$ *is dense in* $\mathcal{M}$ *if for every* $u \in \mathcal{M}$ *there exists a sequence* $\{u_n\} \subset \mathcal{M}$ *such that* $u = \lim_{n\to\infty} u_n$.

Definition 14 *Two metric spaces* $\mathcal{M}_\infty, \mathcal{M}_\in$ *are isometric if there is a 1-1 onto map* $f : \mathcal{M}_\infty \to \mathcal{M}_\in$ *such that* $\rho_2(f(u), f(v)) = \rho_1(u, v)$ *for all* $u, v \in \mathcal{M}_\infty$

Remark: We identify isometric spaces.

Theorem 4 *Given an incomplete metric space* $\mathcal{M}$ *we can extend it to a complete metric space* $\overline{\mathcal{M}}$ *(the completion of* $\mathcal{M}$*) such that 1)* $\mathcal{M}$ *is dense in* $\overline{\mathcal{M}}$. *2) Any two such completions* $\overline{\mathcal{M}}', \overline{\mathcal{M}}''$ *are isometric.*

PROOF: (divided into parts)

1. **Definition 15** *Two Cauchy sequences* $\{u_n\}, \{\tilde{u}_n\}$ *in* $\mathcal{M}$ *are equivalent (*$\{u_n\} \sim \{\tilde{u}_n\}$*) if* $\rho(u_n, \tilde{u}_n) \to 0$ *as* $n \to \infty$.

   Clearly $\sim$ is an equivalence relation, i.e.,

   (a) $\{u_n\} \sim \{u_n\}$, reflexive

   (b) If $\{u_n\} \sim \{v_n\}$ then $\{v_n\} \sim \{u_n\}$, symmetric

   (c) If $\{u_n\} \sim \{v_n\}$ and $\{v_n\} \sim \{w_n\}$ then $\{u_n\} \sim \{v_n\}$. transitive

   Let $\overline{\mathcal{M}}$ be the set of all equivalence classes of Cauchy sequences. An equivalence class $\overline{u}$ consists of all Cauchy sequences equivalent to a given $\{u_n\}$.

2. $\overline{\mathcal{M}}$ is a metric space. Define $\overline{\rho}(\overline{u}, \overline{v}) = \lim_{n\to\infty} \rho(u_n, v_n)$, where $\{u_n\} \in \overline{u}, \{v_n\} \in \overline{v}$.

(a) $\overline{\rho}(\overline{u}, \overline{v})$ exists.

PROOF:

$$\rho(u_n, v_n) \leq \rho(u_n, u_m) + \rho(u_m, v_m) + \rho(v_m, v_n),$$

so

$$\rho(u_n, v_n) - \rho(u_m, v_m) \leq \rho(u_n, u_m) + \rho(v_m, v_n),$$

and

$$|\rho(u_n, v_n) - \rho(u_m, v_m)| \leq \rho(u_n, u_m) + \rho(v_m, v_n) \to 0$$

as $n, m \to \infty$.

(b) $\overline{\rho}(\overline{u}, \overline{v})$ is well defined.

PROOF: Let $\{u_n\}, \{u'_n\} \in \overline{u}, \{v_n\}, \{v'_n\} \in \overline{v}$. Does $\lim_{n\to\infty} \rho(u_n, v_n) = \lim_{n\to\infty} \rho(u'_n, v'_n)$? Yes, because

$$\rho(u_n, v_n) \leq \rho(u_n, u'_n) + \rho(u'_n, v'_n) + \rho(v'_n, v_n),$$

so

$$|\rho(u_n, v_n) - \rho(u'_n, v'_n)| \leq \rho(u_n, u'_n) + \rho(v'_n, v_n) \to 0$$

as $n \to \infty$.

(c) $\overline{\rho}$ is a metric on $\overline{\mathcal{M}}$, i.e.

i. $\overline{\rho}(\overline{u}, \overline{v}) \geq 0$, and $= 0$ if and only if $\overline{u} = \overline{v}$

PROOF: $\overline{\rho}(\overline{u}, \overline{v}) = \lim_{n\to\infty} \rho(u_n, v_n) \geq 0$ and $= 0$ if and only if $\{u_n\} \sim \{v_n\}$, i.e., if and only if $\overline{u} = \overline{v}$.

ii. $\overline{\rho}(\overline{u}, \overline{v}) = \overline{\rho}(\overline{v}, \overline{u})$ obvious

iii. $\overline{\rho}(\overline{u}, \overline{v}) \leq \overline{\rho}(\overline{u}, \overline{w}) + \overline{\rho}(\overline{w}, \overline{v})$ easy

(d) $\mathcal{M}$ is isometric to a metric subset $\overline{\mathcal{S}}$ of $\overline{\mathcal{M}}$.

PROOF: Consider the set $\overline{\mathcal{S}}$ of equivalence classes $\overline{u}$ all of whose Cauchy sequences converge to elements of $\mathcal{M}$. If $\overline{u}$ is such a class then there exists $u \in \mathcal{M}$ such that $\lim_{n\to\infty} u_n = u$ if $\{u_n\} \in \overline{u}$. Note that $u, u, \cdots, u, \cdots \in \overline{u}$ (stationary sequence). The map $u \leftrightarrow \overline{u}$ is a 1-1 map of $\mathcal{M}$ onto $\overline{\mathcal{S}}$. It is an isometry since

$$\overline{\rho}(\overline{u}, \overline{v}) = \lim_{n\to\infty} \rho(u_n, v_n) = \rho(u, v)$$

for $\overline{u}, \overline{v} \in \overline{\mathcal{S}}$, with $\{u_n\} = \{u\} \in \overline{u}, \{v_n\} = \{v\} \in \overline{v}$.

(e) $\mathcal{M}$ is dense in $\overline{\mathcal{M}}$.

PROOF: Let $\overline{u} \in \overline{\mathcal{M}}$, $\{u_n\} \in \overline{u}$. Consider $\overline{s}_k = \{u_k, u_k, \cdots, u_k, \cdots\} \in \overline{\mathcal{S}} = \mathcal{M}$, $k = 1, 2, \cdots$. Then $\overline{\rho}(\overline{u}, \overline{s}_k) = \lim_{n\to\infty} \rho(u_n, u_k)$. But $\{u_n\}$ is Cauchy in $\mathcal{M}$. Therefore, given $\epsilon > 0$, if we choose $k > N(\epsilon)$ we have $\overline{\rho}(\overline{u}, \overline{s}_k) < \epsilon$. Q.E.D.

(f) $\overline{\mathcal{M}}$ is complete.

PROOF: Let $\{\overline{v}_k\}$ be a Cauchy sequence in $\overline{\mathcal{M}}$. For each k choose $\overline{s}_k = \{u_k, u_k, \cdots, u_k, \cdots\} \in \overline{\mathcal{S}} = \mathcal{M}$, such that $\overline{\rho}(\overline{v}_k, \overline{s}_k) < 1/k$, $k = 1, 2, \cdots$. Then

$$\rho(u_j, u_k) = \overline{\rho}(\overline{s}_j, \overline{s}_k) \leq \overline{\rho}(\overline{s}_j, \overline{v}_j) + \overline{\rho}(\overline{v}_j, \overline{v}_k) + \overline{\rho}(\overline{v}_k, \overline{s}_k) \to 0$$

as $j, k \to \infty$. Therefore $\overline{u} = \{u_k\}$ is Cauchy in $\mathcal{M}$. Now

$$\overline{\rho}(\overline{u}, \overline{v}_k) \leq \overline{\rho}(\overline{u}, \overline{s}_k) + \overline{\rho}(\overline{s}_k, \overline{v}_k) \to 0$$

as $k \to \infty$. Therefore $\lim_{k\to\infty} \overline{v}_k = \overline{u}$. Q.E.D.

2.3.1 Completion of a normed linear space

Here $\mathcal{B}$ is a normed linear space with norm $\rho(u, v) = ||u - v||$. We will show how to extend it to a complete normed linear space, called a *Banach Space*.

Definition 16 *Let $\mathcal{S}$ be a subspace of the normed linear space $\mathcal{B}$. $\mathcal{S}$ is a dense subspace of $\mathcal{B}$ if it is a dense subset of $\mathcal{B}$. $\mathcal{S}$ is a closed subspace of $\mathcal{B}$ if every Cauchy sequence $\{u_n\}$ in $\mathcal{S}$ converges to an element of $\mathcal{S}$. (Note: If $\mathcal{B}$ is a Banach space then so is $\mathcal{S}$.)*

Theorem 5 *An incomplete normed linear space $\mathcal{B}$ can be extended to a Banach space $\overline{\mathcal{B}}$ such that $\mathcal{B}$ is a dense subspace of $\overline{\mathcal{B}}$.*

PROOF: By the previous theorem we can extend the metric space $\mathcal{B}$ to a complete metric space $\overline{\mathcal{B}}$ such that $\mathcal{B}$ is dense in $\overline{\mathcal{B}}$.

1. $\overline{\mathcal{B}}$ is a vector space.

(a) $\overline{u}, \overline{v} \in \overline{\mathcal{B}} \longrightarrow \overline{u} + \overline{v} \in \overline{\mathcal{B}}$.

If $\{u_n\} \in \overline{u}, \quad \{v_n\} \in \overline{v}$, define $\overline{u} + \overline{v} = \overline{u+v}$ as the equivalence class containing $\{u_n + v_n\}$. Now $\{u_n + v_n\}$ is Cauchy because $||(u_n + v_n) - (u_m - v_m)|| \leq ||u_n - u_m|| + ||v_n - v_m|| \to 0$ as $n, m \to \infty$. Easy to check that addition is well defined.

(b) $\alpha \in C, \overline{u} \in \overline{\mathcal{B}} \longrightarrow \alpha\overline{u} \in \overline{\mathcal{B}}$.

If $\{u_n\} \in \overline{u}$, define $\alpha\overline{u} \in \overline{\mathcal{B}}$ as the equivalence class containing $\{\alpha u_n\}$, Cauchy because $||\alpha u_n - \alpha u_m|| \leq |\alpha|||u_n - u_m||$.

2. $\overline{\mathcal{B}}$ is a Banach space.

Define the norm $||\overline{u}||'$ on $\overline{\mathcal{B}}$ by $||\overline{u}||' = \overline{\rho}(\overline{u}, \overline{\Theta}) = \lim_{n\to\infty} ||u_n||$ where $\overline{\Theta}$ is the equivalence class containing $\{\Theta, \Theta, \cdots\}$. positivity is easy. Let $\alpha \in C$, $\{u_n\} \in \overline{u}$. Then $||\alpha\overline{u}||' = \overline{\rho}(\alpha\overline{u}, \overline{\Theta}) = \lim_{n\to\infty} ||\alpha u_n|| = |\alpha| \lim_{n\to\infty} ||u_n|| = |\alpha|\overline{\rho}(\overline{u}, \overline{\Theta}) = |\alpha|||\overline{u}||'$.

$||\overline{u} + \overline{v}||' = \overline{\rho}(\overline{u} + \overline{v}, \overline{\Theta}) \leq \overline{\rho}(\overline{u} + \overline{v}, \overline{v}) + \overline{\rho}(\overline{v}, \overline{\Theta}) = ||\overline{u}||' + ||\overline{v}||'$, because $\overline{\rho}(\overline{u} + \overline{v}, \overline{v}) = \lim_{n\to\infty} ||(u_n + v_n) - v_n|| = \lim_{n\to\infty} ||u_n|| = ||\overline{u}||'$. Q.E.D.

2.3.2 Completion of an inner product space

Here $\mathcal{H}$ is an inner product space with inner product (u, v) and norm $\rho(u, v) = ||u - v||$. We will show how to extend it to a complete inner product space, called a *Hilbert Space*.

Theorem 6 *Let $\mathcal{H}$ be an inner product space and $\{u_n\}, \{v_n\}$ convergent sequences in $\mathcal{H}$ with $\lim_{n\to\infty} u_n = u, \lim_{n\to\infty} v_n = v$. Then $\lim_{n\to\infty}(u_n, v_n) = (u, v)$.*

PROOF: Must first show that $||u_n||$ is bounded for all n. $\{u_n\}$ converges $\longrightarrow$ $||u_n|| \leq ||u_n - u|| + ||u|| < \epsilon + ||u||$ for $n > N(\epsilon)$. Set $K = \max\{||u_1||, \cdots, ||u_{N(\epsilon)||}, \epsilon + ||u||\}$. Then $||u_n|| \leq K$ for all n. Then $|(u, v) - (u_n, v_n)| = |(u - u_n, v) + (u_n, v - v_n)| \leq ||u - u_n|| \cdot ||v|| + ||u_n|| \cdot ||v - v_n|| \to 0$ as $n \to \infty$. Q.E.D.

Theorem 7 *Let $\mathcal{H}$ be an incomplete inner product space. We can extend $\mathcal{H}$ to a Hilbert space $\overline{\mathcal{H}}$ such that $\mathcal{H}$ is a dense subspace of $\overline{\mathcal{H}}$.*

PROOF: $\mathcal{H}$ is a normed linear space with norm $||u|| = \sqrt{(u, u)}$. Therefore we can extend $\mathcal{H}$ to a Banach space $\overline{\mathcal{H}}$ such that $\mathcal{H}$ is dense in $\overline{\mathcal{H}}$. Claim that $\overline{\mathcal{H}}$

is a Hilbert space. Let $\overline{u}, \overline{v} \in \overline{\mathcal{H}}$ and let $\{u_n\}, \{\tilde{u}_n\} \in \overline{u}$, $\{v_n\}, \{\tilde{v}_n\} \in \overline{v}$. We define an inner product on $\overline{\mathcal{H}}$ by $(\overline{u}, \overline{v})' = \lim_{n\to\infty}(u_n, v_n)$. The limit exists since $|(u_n, v_n) - (u_m, v_m)| = |(u_m, v_n - v_m) + (u_n - u_m, v_m) + (u_n - u_m, v_n - v_m)| \leq ||u_m|| \cdot ||v_n - v_m|| + ||u_n - u_m|| \cdot ||v_m|| + ||u_n - u_m|| \cdot ||v_n - v_m|| \to 0$ as $n, m \to \infty$. The limit is unique because $|(u_n, v_n) - (\tilde{u}_n, \tilde{v}_n)| \to 0$ as $n, m \to \infty$. can easily verify that $(\cdot, \cdot)'$ is an inner product on $\overline{\mathcal{H}}$ and $|| \cdot ||' = \sqrt{(\cdot, \cdot)'}$. Q.E.D.

2.4 Hilbert spaces, L^2 and ℓ^2

A Hilbert space is an inner product space for which every Cauchy sequence in the norm converges to an element of the space.

EXAMPLE: ℓ^2

The elements take the form

$$u = (\cdots, \alpha_{-1}, \alpha_0, \alpha_1, \cdots), \quad \alpha_i \in C$$

such that $\sum_{i=-\infty}^{\infty} |\alpha_i|^2 < \infty$. For

$$v = (\cdots, \beta_{-1}, \beta_0, \beta_1, \cdots) \in \ell^2,$$

we define vector addition and scalar multiplication by

$$u + v = (\cdots, \alpha_{-1} + \beta_{-1}, \alpha_0 + \beta_0, \alpha_1 + \beta_1, \cdots)$$

and

$$\alpha u = (\cdots, \alpha\alpha_{-1}, \alpha\alpha_0, \alpha\alpha_1, \cdots).$$

The zero vector is $\Theta = (\cdots, 0, 0, 0, \cdots)$ and the inner product is defined by $(u.v) = \sum_{i=-\infty}^{\infty} \alpha_i \bar{\beta}_i$. We have to verify that these definitions make sense. Note that $2|ab| \leq |a|^2 + |b|^2$ for any $a, b \in C$. The inner product is well defined because $|(u, v)| \leq \sum_{i=-\infty}^{\infty} |\alpha_i \bar{\beta}_i| \leq \frac{1}{2}(\sum_{i=-\infty}^{\infty} |\alpha_i|^2 + \sum_{i=-\infty}^{\infty} |\beta_i|^2) < \infty$. Note that $|\alpha_i + \beta_i|^2 \leq |\alpha_i|^2 + 2|\alpha_i| \cdot |\beta_i| + |\beta_i|^2 \leq 2(|\alpha_i|^2 + |\beta_i|^2)$. Thus if $u, v \in \ell^2$ we have $||u + v||^2 \leq 2||u||^2 + 2||v||^2 < \infty$, so $u + v \in \ell^2$.

Theorem 8 *ℓ^2 is a Hilbert space.*

PROOF: We have to show that ℓ^2 is complete. Let $\{u_n\}$ be Cauchy in ℓ^2,

$$u_n = (\cdots, \alpha_{-1}^{(n)}, \alpha_0^{(n)}, \alpha_1^{(n)}, \cdots).$$

Thus, given any $\epsilon > 0$ there exists an integer $N(\epsilon)$ such that $||u_n - u_m|| < \epsilon$ whenever $n, m > N(\epsilon)$. Thus

$$\sum_{i=-\infty}^{\infty} |\alpha_i^{(n)} - \alpha_i^{(m)}|^2 < \epsilon^2. \tag{2.1}$$

Hence, for fixed i we have $|\alpha_i^{(n)} - \alpha_i^{(m)}| < \epsilon$. This means that for each i, $\{\alpha_i^{(n)}\}$ is a Cauchy sequence in C. Since C is complete, there exists $\alpha_i \in C$ such that $\lim_{n\to\infty} \alpha_i^{(n)} = \alpha_i$ for all integers i. Now set $u = (\cdots, \alpha_{-1}, \alpha_0, \alpha_1, \cdots)$. Claim that $u \in \ell^2$ and $\lim_{n\to\infty} u_n = u$. It follows from (2.1) that for any fixed k, $\sum_{i=-\infty}^{k} |\alpha_i^{(n)} - \alpha_i^{(m)}|^2 < \epsilon^2$ for $n, m > N(\epsilon)$. Now let $m \to \infty$ and get $\sum_{i=-\infty}^{k} |\alpha_i^{(n)} - \alpha_i|^2 \leq \epsilon^2$ for all k and for $n > N(\epsilon)$. Next let $k \to \infty$ and get $\sum_{i=-\infty}^{\infty} |\alpha_i^{(n)} - \alpha_i|^2 \leq \epsilon^2$ for $n > N(\epsilon)$. This implies

$$||u_n - u|| < \epsilon \tag{2.2}$$

for $n > N(\epsilon)$. Thus, $u_n - u \in \ell^2$ for $n > N(\epsilon)$, so $u = (u - u_n) + u_n \in \ell^2$. Finally, (2.2) implies that $\lim_{n\to\infty} u_n = u$. Q.E.D.

EXAMPLE: $L^2[a, b]$

Recall that $C_2(a, b)$ is the set of all complex-valued functions $u(t)$ continuous on the open interval (a, b) of the real line, such that $\int_a^b |u(t)|^2\, dt < \infty$, (Riemann integral). We define an inner product by

$$(u, v) = \int_a^b u(t)\overline{v}(t)\, dt, \qquad u, v \in C_2^{(n)}(a, b).$$

We verify that this is an inner product space. First, from the inequality $|u(x) + v(x)|^2 \leq 2|u(x)|^2 + 2|v(x)|^2$ we have $||u + v||^2 \leq 2||u||^2 + 2||v||^2$, so if $u, v \in C_2(a, b)$ then $u + v \in C_2(a, b)$. Second, $|u(x)\overline{v}(x)| \leq \frac{1}{2}(|u(x)|^2 + |v(x)|^2)$, so $|(u, v)| \leq \int_a^b |u(t)\overline{v}(t)|\, dt \leq \frac{1}{2}(||u||^2 + ||v||^2) < \infty$ and the inner product is well defined.

Now $C_2(a, b)$ is not complete, but it is dense in a Hilbert space $\overline{C}_2(a, b) = \overline{L}_0^2[a, b] = L^2[a, b]$ In most of this course we will normalize to the case $a = 0, b = 2\pi$. We will show that the functions $e_n(t) = e^{int}/\sqrt{2\pi}$, $n = 0, \pm 1, \pm 2, \cdots$ form a basis for $L^2[0, 2\pi]$. This is a countable (rather than a continuum) basis. Hilbert spaces with countable bases are called *separable*, and we will be concerned only with separable Hilbert spaces in this course.

2.4.1 The Riemann integral and the Lebesgue integral

Recall that $S^1(J)$ is the normed linear space space of all real or complex-valued step functions on the (bounded or unbounded) interval J on the real line. s is a *step function* on J if there are a finite number of non-intersecting bounded intervals $J_1, \cdots, J_m$ and numbers $c_1, \cdots, c_m$ such that $s(t) = c_k$ for $t \in J_k$, $k = 1, \cdots, m$ and $s(t) = 0$ for $t \in J - \cup_{k=1}^m J_k$. The integral of a step function is the $\int_J s(t)dt \equiv \sum_{k=1}^m c_k \ell(J_k)$ where $\ell(J_k) =$ length of J_k = $b - a$ if $J_k = [a, b]$ or $[a, b)$, or $(a, b]$ or (a, b). The norm is defined by $||s|| = \int_J |s(t)|dt$. We identify $s_1, s_2 \in S(J)$, $s_1 \sim s_2$ if $s_1(t) = s_2(t)$ except at a finite number of points. (This is needed to satisfy property 1. of the norm.) We let $S^1(J)$ be the space of equivalence classes of step functions in $S(J)$. Then $S^1(J)$ is a normed linear space with norm $|| \cdot ||$.

The space of *Lebesgue integrable functions* on J, ($L^1(J)$) is the completion of $S^1(J)$ in this norm. $L^1(J)$ is a Banach space. Every element u of $L^1(J)$ is an equivalence class of Cauchy sequences of step functions $\{s_n\}$, $\int_J |s_j - s_k|dt \to 0$ as $j, k \to \infty$. (Recall $\{s'_n\} \sim \{s_n\}$ if $\int_J |s'_k - s_n|dt \to 0$ as $n \to \infty$.

It is beyond the scope of this course to prove it, but, in fact, we can associate equivalence classes of functions $f(t)$ on J with each equivalence class of step functions $\{s_n\}$. The Lebesgue integral of f is defined by

$$\int_{J \text{ Lebesgue}} f(t)dt = \lim_{n\to\infty} \int_J s_n(t)dt,$$

and its norm by

$$||f|| = \int_{J \text{ Lebesgue}} |f(t)|dt = \lim_{n\to\infty} \int_J |s_n(t)|dt.$$

How does this definition relate to Riemann integrable functions? To see this we take $J = [a, b]$, a closed bounded interval, and let $f(t)$ be a real bounded function on $[a, b]$. Recall that we have already defined the integral of a step function.

Definition 17 *f is Riemann integrable on $[a, b]$ if for every $\epsilon > 0$ there exist step functions $r, s \in S[a, b]$ such that $r(t) \le f(t) \le s(t)$ for all $t \in [a, b]$, and $0 \le \int_a^b (s - r)dt < \epsilon$.*

EXAMPLE. Divide $[a, b]$ by a grid of n points $a = t_0 < t_1 < \cdots < t_n = b$ such that $t_j - t_{j-1} = (b - a)/n$, $j = 1, \cdots, n$. Let $M_j = \sup_{t\in[t_{j-1},t_j]} f(t)$, $m_j = \inf_{t\in[t_{j-1},t_j]} f(t)$ and set

$$s_n(t) = \{ \begin{array}{ll} M_j & t \in [t_{j-1}, t_j) \\ 0 & t \notin [a, b) \end{array}$$

$$r_n(t) = \{ \begin{array}{ll} m_j & t \in [t_{j-1}, t_j) \\ 0 & t \notin [a,b) \end{array}$$

$\int_a^b s_n(t)dt$ is an *upper Darboux sum.* $\int_a^b r_n(t)dt$ is a *lower Darboux sum.* If f is Riemann integrable then the sequences of step functions $\{r_n\}, \{s_n\}$ satisfy $r_n \le f \le s_n$ on $[a,b]$, for $n = 1, 2, \cdots$ and $\int_a^b (s_n - r_n)dt \to 0$ as $n \to \infty$. The Riemann integral is defined by

$$\int_{a\text{Riemann}}^b f\ dt = \lim_{n\to\infty} \int s_n\ dt = \lim_{n\to\infty} \int r_n\ dt =$$

$$\inf_{\text{upper Darboux sums}} \int s\ dt = \sup_{\text{lower Darboux sums}} \int t\ dt.$$

Note that

$$\sum_{j=1}^n M_j(t_j - t_{j-1}) \ge \int_{a\ \text{Riemann}}^b f\ dt \ge \sum_{j=1}^n M_j(t_j - t_{j-1}).$$

Note also that

$$r_j - r_k \le s_k - r_k, \qquad r_k - r_j \le s_j - r_j$$

because every “upper” function is $\ge$ every “lower” function. Thus

$$\int |r_j - r_k|dt \le \int (s_k - r_k)dt + \int (s_j - r_j)dt \to 0$$

as $j, k \to \infty$. Thus $\{r_n\}$ and similarly $\{s_n\}$ are Cauchy sequences in the norm, equivalent because $\lim_{n\to\infty} \int (s_n - r_n)dt = 0$.

Theorem 9 *If f is Riemann integrable on $J = [a,b]$ then it is also Lebesgue integrable and*

$$\int_{J\ \text{Riemann}} f(t)dt = \int_{J\ \text{Lebesgue}} f(t)dt = \lim_{n\to\infty} \int_J s_n(t)dt$$

.

The following is a simple example to show that the space of Riemann integrable functions isn’t complete. Consider the closed interval $J = [0,1]$ and let $r_1, r_2, \cdots$ be an enumeration of the rational numbers in $[0,1]$. Define the sequence of step functions $\{s_n\}$ by

$$s_n(t) = \{ \begin{array}{ll} 1 & t = r_1, r_2, \cdots, r_n \\ 0 & \text{otherwise.} \end{array}$$

Note that

- $s_1(t) \leq s_2(t) \leq \cdots$ for all $t \in [0,1]$.
- s_n is a step function.
- The pointwise limit

$$f(t) = \lim_{n\to\infty} s_n(t) = \{ \begin{array}{ll} 1 & \text{if } t \text{ is rational} \\ 0 & \text{otherwise.} \end{array}$$

- $\{s_n\}$ is Cauchy in the norm. Indeed $\int_0^1 |s_j - s_k|dt = 0$ for all $j, k = 1, 2, \cdots$.
- f is Lebesgue integrable with $\int_0^1{}_{\text{Lebesgue}} f(t)dt = \lim_{n\to\infty} \int_0^1 s_n(t)dt = 0$.
- f is not Riemann integrable because *every* upper Darboux sum for f is 1 and every lower Darboux sum is 0. Can't make $1 - 0 < \epsilon$ for $\epsilon < 1$.

Recall that $S^2(J)$ is the space of all real or complex-valued step functions on the (bounded or unbounded) interval J on the real line with real inner product b $(s_1, s_2) = \int_J s_1(t)\bar{s}_2(t)dt$. We identify $s_1, s_2 \in S(J)$, $s_1 \sim s_2$ if $s_1(t) = s_2(t)$ except at a finite number of points. (This is needed to satisfy property 4. of the inner product.) Now we let $S^2(J)$ be the space of equivalence classes of step functions in $S(J)$. Then $S^2(J)$ is an inner product space with norm $||s||^2 = \int_J |s(t)|^2 dt$.

The space of *Lebesgue square-integrable functions* on J, ($L^2(J)$) is the completion of $S^2(J)$ in this norm. $L^2(J)$ is a Hilbert space. Every element u of $L^2(J)$ is an equivalence class of Cauchy sequences of step functions $\{s_n\}$, $\int_J |s_j - s_k|^2 dt \to 0$ as $j, k \to \infty$. (Recall $\{s'_n\} \sim \{s_n\}$ if $\int_J |s'_k - s_n|^2 dt \to 0$ as $n \to \infty$.

It is beyond the scope of this course to prove it, but, in fact, we can associate equivalence classes of functions $f(t)$ on J with each equivalence class of step functions $\{s_n\}$. The Lebesgue integral of $f_1, f_2 \in L^2(J)$ is defined by $(f_1, f_2) = \int_J{}_{\text{Lebesgue}} f_1(t)f_2\, dt = \lim_{n\to\infty} \int_J s_n^{(1)}(t)s_n^{(2)}(t)dt$.

How does this definition relate to Riemann square integrable functions? In a manner similar to our treatment of $L^1(J)$ one can show that if the function f is Riemann square integrable on J, then it is Cauchy square integrable and $\int_J{}_{\text{Lebesgue}} |f(t)|^2 dt = \int_J{}_{\text{Riemann}} |f(t)|^2 dt$.

2.5 Orthogonal projections, Gram-Schmidt orthogonalization

2.5.1 Orthogonality, Orthonormal bases

Definition 18 *Two vectors u, v in an inner product space $\mathcal{H}$ are called orthogonal, $u \perp v$, if $(u, v) = 0$. Similarly, two sets $\mathcal{M}, \mathcal{N} \subset \mathcal{H}$ are orthogonal, $\mathcal{M} \perp \mathcal{N}$, if $(u, v) = 0$ for all $u \in \mathcal{M}$, $v \in \mathcal{N}$.*

Definition 19 *Let $\mathcal{S}$ be a nonempty subset of the inner product space $\mathcal{H}$. We define $\mathcal{S}^\perp$ by $\mathcal{S}^\perp = \{u \in \mathcal{H} : u \perp \mathcal{S}\}$*

Lemma 3 *$\mathcal{S}^\perp$ is closed subspace of $\mathcal{H}$ in the sense that if $\{u_n\}$ is a Cauchy sequence in $\mathcal{S}^\perp$ and $u_n \to u \in \mathcal{H}$ as $n \to \infty$ then $u \in \mathcal{S}^\perp$.*

PROOF:

1. $\mathcal{S}^\perp$ is a subspace. Let $u, v \in \mathcal{S}^\perp$, $\alpha, \beta \in C$, Then $(\alpha u + \beta v, w) = \alpha(u, w) + \beta(v, w) = 0$ for all $w \in \mathcal{S}$, so $\alpha u + \beta v \in \mathcal{S}^\perp$.

2. $\mathcal{S}^\perp$ is closed. Suppose $\{u_n\} \subset \mathcal{S}^\perp$, $\lim_{n\to\infty} u_n = u \in \mathcal{H}$. Then $(u, v) = (\lim_{n\to\infty} u_n, v) = \lim_{n\to\infty}(u_n, v) = 0$ for all $v \in \mathcal{S} \Longrightarrow u \in \mathcal{S}^\perp$. Q.E.D.

2.5.2 Orthonormal bases for finite-dimensional inner product spaces

Let $\mathcal{H}$ be an n-dimensional inner product space, (say $\mathcal{H}_n$). A vector $u \in \mathcal{H}$ is a *unit vector* if $||u|| = 1$. The elements of a finite subset $\{u_1, \cdots, u_k\} \subset \mathcal{H}$ are *mutually orthogonal* if $u_i \perp u_j$ for $i \neq j$. The finite subset $\{u_1, \cdots, u_k\} \subset \mathcal{H}$ is *orthonormal* (ON) if $u_i \perp u_j$ for $i \neq j$, and $||u_i|| = 1$. Orthonormal bases for $\mathcal{H}$ are especially convenient because the expansion coefficients of any vector in terms of the basis can be calculated easily from the inner product.

Theorem 10 *Let $\{u_1, \cdots, u_n\}$ be an ON basis for $\mathcal{H}$. If $u \in \mathcal{H}$ then*

$$u = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n$$

where $\alpha_i = (u, u_i)$, $i = 1, \cdots, n$.

PROOF: $(u, u_i) = (\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n, u_i) = \alpha_1$. Q.E.D.

Example 1 *Consider $\mathcal{H}_3$. The set $e_1 = (1,0,0), e_2 = (0,1,0), e_3 = (0,0,1)$ is an ON basis. The set $u_1 = (1,0,0), u_2 = (1,1,0), u_3 = (1,1,1)$ is a basis, but not ON. The set $v_1 = (1,0,0), v_2 = (0,2,0), v_3 = (0,0,3)$ is an orthogonal basis, but not ON.*

The following are very familiar results from geometry, where the inner product is the dot product, but apply generally and are easy to prove:

Corollary 1 *For $u, v \in \mathcal{H}$:*

- $(u, v) = (\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n, \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_n u_n) = \sum_{i=1}^{n} (u, u_i)(u_i, v)$
- $||u||^2 = \sum_{i=1}^{n} |(u, u_i)|^2$ *Parseval's equality.*

Lemma 4 *If $u \perp v$ then $||u + v||^2 = ||u||^2 + ||v||^2$ Pythagorean Theorem*

Lemma 5 *For any $u, v \in \mathcal{H}$ we have $||u + v||^2 + ||u - v||^2 = 2||u||^2 + 2||v||^2$.*

Lemma 6 *If u, v belong to the real inner product space $\mathcal{H}$ then $||u + v||^2 = ||u||^2 + ||v||^2 + 2(u, v)$. Law of Cosines.*

Note: The preceding lemmas are obviously true for any inner product space, finite-dimensional or not.

Does every n-dimensional inner product space have an ON basis? Yes!

Recall that $[u_1, u_2, \cdots, u_m]$ is the subspace of $\mathcal{H}$ spanned by all linear combinations of the vectors $u_1, u_2, \cdots, u_m$.

Theorem 11 *(Gram-Schmidt) let $\{u_1, u_2, \cdots, u_n\}$ be an (ordered) basis for the inner product space $\mathcal{H}$. There exists an ON basis $\{e_1, e_2, \cdots, e_n\}$ for $\mathcal{H}$ such that*

$$[u_1, u_2, \cdots, u_m] = [e_1, e_2, \cdots, e_m]$$

for each $m = 1, 2, \cdots, n$.

PROOF: Define e_1 by $e_i = u_1/||u_1||$ This implies $||e_1|| = 1$ and $[u_1] = [e_1]$. Now set $f_2 = u_2 - \alpha e_1 \neq \Theta$. We determine the constant α by requiring that $(f_2, e_1) = 0$ But $(f_2, e_1) = (u_2, e_1) - \alpha$ so $\alpha = (u_2, e_1)$. Now define e_2 by $e_2 = f_2/||f_2||$. At this point we have $(e_i, e_j) = \delta_{ij}$ for $1 \leq i, j \leq 2$ and $[u_1, u_2] = [e_1, e_2]$.

We proceed by induction. Assume we have constructed an ON set $\{e_1, \cdots, e_m\}$ such that $[e_1, \cdots, e_k] = [u_1, \cdots, u_k]$ for $k = 1, 2, \cdots, m$. Set $f_{m+1} = u_{m+1} - \alpha_1 e_1 - \alpha_2 e_2 - \cdots - \alpha_m e_m \neq 0$. Determine the constants α_i by the requirement $(f_{m+1}, e_i) = 0 = (u_{m+1}, e_i) - \alpha_i$, $1 \le i \le m$. Set $e_{m+1} = f_{m+1}/||f_{m+1}||$. Then $\{e_1, \cdots, e_{m+1}\}$ is ON. Q.E.D.

Let $\mathcal{W}$ be a subspace of $\mathcal{H}$ and let $\{e_1, e_2, \cdots, e_m\}$ be an ON basis for $\mathcal{W}$. let $u \in \mathcal{H}$. We say that the vector $u' = \sum_{i=1}^m (u, e_i)e_i \in \mathcal{W}$ is the *projection of* u *on* $\mathcal{W}$.

Theorem 12 *If $u \in \mathcal{H}$ there exist unique vectors $u' \in \mathcal{W}$, $u'' \in \mathcal{W}^\perp$ such that $u = u' + u''$.*

PROOF:

1. Existence: Let $\{e_1, e_2, \cdots, e_m\}$ be an ON basis for $\mathcal{W}$, set $u' = \sum_{i=1}^m (u, e_i)e_i \in \mathcal{W}$ and $u'' = u - u'$. Now $(u'', e_i) = (u, e_i) - (u, e_i) = 0$, $1 \le i \le m$, so $(u'', v) = 0$ for all $v \in \mathcal{W}$. Thus $u'' \in \mathcal{W}^\perp$.

2. Uniqueness: Suppose $u = u' + u'' = v" + v''$ where $u', v' \in \mathcal{W}$, $u'', v'' \in \mathcal{W}^\perp$. Then $u' - v' = v'' - u'' \in \mathcal{W} \cap \mathcal{W}^\perp \Longrightarrow (u' - v', u' - v') = 0 = ||u' - v'||^2 \Longrightarrow u' = v', u'' = v''$. Q.E.D.

Corollary 2 *Bessel's Inequality. Let $\{e_1, \cdots, e_m\}$ be an ON set in $\mathcal{H}$. if $u \in \mathcal{H}$ then $||u||^2 \ge \sum_{i=1}^m |(u, e_i)|^2$.*

PROOF: Set $W = [e_1, \cdots, e_m]$. Then $u = u' + u''$ where $u' \in \mathcal{W}$, $u'' \in \mathcal{W}^\perp$, and $u' = \sum_{i=1}^m (u, e_i)e_i$. Therefore $||u||^2 = (u' + u'', u' + u'') = ||u'||^2 + ||u''||^2 \ge ||u'||^2 = (u', u') + \sum_{i=1}^m |(u, e_i)|^2$. Q.E.D.

Note that this inequality holds even if m is infinite.

The projection of $u \in \mathcal{H}$ onto the subspace $\mathcal{W}$ has invariant meaning, i.e., it is basis independent. Also, it solves an important minimization problem: u' is the vector in $\mathcal{W}$ that is closest to u.

Theorem 13 $\min_{v \in \mathcal{W}} ||u - v|| = ||u - u'||$ *and the minimum is achieved if and only if $v = u'$.*

PROOF: let $v \in \mathcal{W}$ and let $\{e_1, e_2, \cdots, e_m\}$ be an ON basis for $\mathcal{W}$. Then $v = \sum_{i=1}^m \alpha_i e_i$ for $\alpha_i = (v, e_i)$ and $||u - v|| = ||u - \sum_{i=1}^m \alpha_i e_i||^2 = (u - \sum_{i=1}^m \alpha_i e_i, u - \sum_{i=1}^m \alpha_i e_i) = ||u||^2 - \sum_{i=1}^m \bar{\alpha}_i (u, e_i) - \sum_{i=1}^m \alpha_i (e_i, u) + \sum_{i=1}^m |\alpha_i|^2 = ||u - \sum_{i=1}^m (u, e_i)e_i||^2 + \sum_{i=1}^m |(u, e_i) - \alpha_i|^2 \ge ||u - u'||^2$. Equality is obtained if and only if $\alpha_i = (u, e_i)$, for $1 \le i \le m$. Q.E.D.

2.5.3 Orthonormal systems in an infinite-dimensional separable Hilbert space

Let $\mathcal{H}$ be a separable Hilbert space. (We have in mind spaces such as ℓ^2 and $L^2[0, 2\pi]$.)

The idea of an orthogonal projection extends to infinite-dimensional inner product spaces, but here there is a problem. If the infinite-dimensional subspace $\mathcal{W}$ of $\mathcal{H}$ isn't closed, the concept may not make sense.

For example, let $\mathcal{H} = \ell^2$ and let $\mathcal{W}$ be the subspace elements of the form $(\cdots, \alpha_{-1}, \alpha_0, \alpha_1, \cdots)$ such that $\alpha_i = 0$ for $i = 1, 0, -1, -2, \cdots$ and there are a *finite* number of nonzero components α_i for $i \geq 2$. Choose $u = (\cdots, \beta_{-1}, \beta_0, \beta_1, \cdots)$ such that $\beta_i = 0$ for $i = 0, -1, -2, \cdots$ and $\beta_n = 1/n$ for $n = 1, 2, \cdots$. Then $u \in \ell^2$ but the projection of u on $\mathcal{W}$ is undefined. If $\mathcal{W}$ is closed, however, i.e., if every Cauchy sequence $\{u_n\}$ in $\mathcal{W}$ converges to an element of $\mathcal{W}$, the problem disappears.

Theorem 14 *Let $\mathcal{W}$ be a closed subspace of the inner product space $\mathcal{H}$ and let $u \in H$. Set $d = \inf_{v\in\mathcal{W}} ||u - v||$. Then there exists a unique $\bar{u} \in \mathcal{W}$ such that $||u - \bar{u}|| = d$, ($\bar{u}$ is called the projection of u on $\mathcal{W}$.) Furthermore $u - \bar{u} \perp \mathcal{W}$ and this characterizes $\bar{u}$.*

PROOF: Clearly there exists a sequence $\{v_n\} \in \mathcal{W}$ such that $||u - v_n|| = d_n$ with $\lim_{n\to\infty} d_n = d$. We will show that $\{v_n\}$ is Cauchy. Using Lemma 5 (which obviously holds for infinite dimensional spaces), we have the equality

$$||(u - v_n) + (u - v_m)||^2 + ||(u - v_n) - (u - v_m)||^2 = 2||u - v_n||^2 + 2||u - v_m||^2$$

or

$$4||u - \frac{1}{2}(v_n + v_m)||^2 + ||v_m - v_n||^2 = 2||u - v_n||^2 + 2||u - v_m||^2.$$

Since $\frac{1}{2}(v_n + v_m) \in \mathcal{W}$ we have the inequality

$$4d^2 + ||v_m - v_n||^2 \leq 4||u - \frac{1}{2}(v_n + v_m)||^2 + ||v_m - v_n||^2 = 2||u - v_n||^2 + 2||u - v_m||^2 = 2(d_n^2 + d_m^2),$$

so

$$||v_m - v_n||^2 \leq 2(d_n^2 - d^2) + 2(d_m^2 - d^2) \to 0,$$

as $n, m \to \infty$. Thus $\{v_n\}$ is Cauchy in $\mathcal{W}$.

Since $\mathcal{W}$ is closed, there exists $\bar{u} \in \mathcal{W}$ such that $\lim_{n\to\infty} v_n = \bar{u}$. Also, $||u - \bar{u}|| = ||u - \lim_{n\to\infty} v_n|| = \lim_{n\to\infty} ||u - v_n|| = \lim_{n\to\infty} d_n = d$. Furthermore, for any $v \in \mathcal{W}$, $(u - \bar{u}, v) = \lim_{n\to\infty}(u - v_n, v) = 0 \Longrightarrow u - \bar{u} \perp \mathcal{W}$.

Conversely, if $u - \bar{u} \perp \mathcal{W}$ and $v \in \mathcal{W}$ then $||u - v||^2 = ||(u - \bar{u}) + (\bar{u} - v)||^2 = ||u - \bar{u}||^2 + ||\bar{u} - v||^2 = d^2 + ||\bar{u} - v||^2$. Therefore $||u - v||^2 \geq d^2$ and $= d^2$ if and only if $\bar{u} = v$. Thus $\bar{u}$ is unique. Q.E.D.

Corollary 3 *Let $\mathcal{W}$ be a closed subspace of the Hilbert space $\mathcal{H}$ and let $u \in \mathcal{H}$. Then there exist unique vectors $\bar{u} \in \mathcal{W}$, $\bar{v} \in \mathcal{W}^\perp$, such that $u = \bar{u} + \bar{v}$. We write $\mathcal{H} = \mathcal{W} \oplus \mathcal{W}^\perp$.*

Corollary 4 *A subspace $\mathcal{M} \subseteq \mathcal{H}$ is dense in $\mathcal{H}$ if and only if $u \perp \mathcal{M}$ for $u \in \mathcal{H}$ implies $u = \Theta$.*

PROOF: $\mathcal{M}$ dense in $\mathcal{H} \Longrightarrow \overline{\mathcal{M}} = \mathcal{H}$. Suppose $u \perp \mathcal{M}$. Then there exists a sequence $\{u_n\}$ in $\mathcal{M}$ such that $\lim_{n\to\infty} u_n = u$ and $(u, u_n) = 0$ for all n. Thus $(u, u) = \lim_{n\to\infty}(u, u_n) = 0 \Longrightarrow u = \Theta$.

Conversely, suppose $u \perp \mathcal{M} \Longrightarrow u = \Theta$. If $\mathcal{M}$ isn't dense in $\mathcal{H}$ then $\bar{\mathcal{M}} \neq \mathcal{H} \Longrightarrow$ there is a $u \in \mathcal{H}$ such that $u \neq \bar{\mathcal{M}}$. Therefore there exists a $\bar{u} \in \bar{\mathcal{M}}$ such that $v = u - \bar{u} \neq \Theta$ belongs to $\bar{\mathcal{M}}^\perp \Longrightarrow v \perp \mathcal{M}$. Impossible! Q.E.D.

Now we are ready to study ON systems on an infinite-dimensional (but separable) Hilbert space $\mathcal{H}$ If $\{v_n\}$ is a sequence in $\mathcal{H}$, we say that $\sum_{n=1}^{\infty} v_n = v \in \mathcal{H}$ if the partial sums $\sum_{n=1}^{k} v_n = u_k$ form a Cauchy sequence and $\lim_{k\to\infty} u_k = v$. This is called *convergence in the mean* or *convergence in the norm*, as distinguished from pointwise convergence of functions. (For Hilbert spaces of functions, such as $L^2[0, 2\pi]$ we need to distinguish this mean convergence from pointwise or uniform convergence.

The following results are just slight extensions of results that we have proved for ON sets in finite-dimensional inner-product spaces. The sequence $u_1, u_2, \cdots \in \mathcal{H}$ is *orthonormal* (ON) if $(u_i, u_j) = \delta_{ij}$. (Note that an ON sequence need not be a basis for $\mathcal{H}$.) Given $u \in \mathcal{H}$, the numbers $\alpha_j = (u, u_j)$ are the *Fourier coefficients* of u with respect to this sequence.

Lemma 7 $u = \sum_{n=1}^{\infty} \alpha_n u_n \Longrightarrow \alpha_n = (u, u_n)$.

Given a fixed ON system $\{u_n\}$, a positive integer N and $u \in \mathcal{H}$ the projection theorem tells us that we can minimize the "error" $||u - \sum_{n=1}^{N} \alpha_n u_n||$ of approximating u by choosing $\alpha_n = (u, u_n)$, i.e., as the Fourier coefficients. Moreover,

Corollary 5 $\sum_{n=1}^{N} |(u, u_n)|^2 \leq ||u||^2$ *for any* N.

Corollary 6 $\sum_{n=1}^{\infty} |(u, u_n)|^2 \leq ||u||^2$, *Bessel's inequality.*

Theorem 15 *Given the ON system* $\{u_n\} \in \mathcal{H}$*, then* $\sum_{n=1}^{\infty} \beta_n u_n$ *converges in the norm if and only if* $\sum_{n=1}^{\infty} |\beta_n|^2 < \infty$.

PROOF: Let $v_k = \sum_{n=1}^{k} \beta_n u_n$. $\sum_{n=1}^{\infty} \beta_n u_n$ converges if and only if $\{v_k\}$ is Cauchy in $\mathcal{H}$. For $k \geq \ell$,

$$||v_k - v_\ell||^2 = ||\sum_{n=\ell+1}^{k} \beta_n u_n||^2 = \sum_{n=\ell+1}^{k} |\beta_n|^2. \tag{2.3}$$

Set $t_k = \sum_{n=1}^{k} |\beta_n|^2$. Then (2.3)$\Longrightarrow$ $\{v_k\}$ is Cauchy in $\mathcal{H}$ if and only if $\{t_k\}$ is Cauchy, if and only if $\sum_{n=1}^{\infty} |\beta_n|^2 < \infty$. Q.E.D.

Definition 20 *A subset* $\mathcal{K}$ *of* $\mathcal{H}$ *is complete if for every* $u \in \mathcal{H}$ *and* $\epsilon > 0$ *there are elements* $u_1, u_2, \cdots, u_N \in \mathcal{K}$ *and* $\alpha_1, \cdots, \alpha_N \in C$ *such that* $||u - \sum_{n=1}^{N} \alpha_n u_n|| < \epsilon$*, i.e., if the subspace* $\tilde{\mathcal{K}}$ *formed by taking all finite linear combinations of elements of* $\mathcal{K}$ *is dense in* $\mathcal{H}$.

Theorem 16 *The following are equivalent for any ON sequence* $\{u_n\}$ *in* $\mathcal{H}$.

1. $\{u_n\}$ *is complete (*$\{u_n\}$ *is an ON basis for* $\mathcal{H}$*.)*

2. *Every* $u \in \mathcal{H}$ *can be written uniquely in the form* $u = \sum_{n=1}^{\infty} \alpha_n u_n$, $\alpha_n = (u, u_n)$.

3. *For every* $u \in \mathcal{H}$, $||u||^2 = \sum_{n=1}^{\infty} |(u, u_n)|^2$. *Parseval's equality*

4. *If* $u \perp \{u_n\}$ *then* $u = \Theta$.

PROOF:

1. 1.$\Longrightarrow$ 2. $\{u_n\}$ complete $\Longrightarrow$ given $u \in \mathcal{H}$ and $\epsilon > 0$ there is an integer N and constants $\{\alpha_n\}$ such that $||u - \sum_{n=1}^{N} \alpha_n u_n|| < \epsilon \Longrightarrow ||u - \sum_{n=1}^{k} \alpha_n u_n|| < \epsilon$ for all $k \geq N$. Clearly $\sum_{n=1}^{\infty}(u, u_n)u_n \in \mathcal{H}$ since $\sum_{n=1}^{\infty} |(u, u_n)|^2 \leq ||u||^2 < \infty$. Therefore $u = \sum_{n=1}^{\infty}(u, u_n)u_n$. Uniqueness obvious.

2. 2.$\Longrightarrow$ 3. Suppose $u = \sum_{n=1}^{\infty} \alpha_n u_n$, $\alpha_n = (u, u_n)$. Therefore, $||u - \sum_{n=1}^{k} \alpha_n u_n||^2 = ||u||^2 - \sum_{n=1}^{k} |(u, u_n)|^2 \to 0$ as $k \to \infty$. Hence $||u||^2 = \sum_{n=1}^{\infty} |(u, u_n)|^2$.

3. 3.$\Longrightarrow$ 4. Suppose $u \perp \{u_n\}$. Then $||u||^2 = \sum_{n=1}^{\infty} |(u, u_n)|^2 = 0$ so $u = \Theta$.

4. 4.$\Longrightarrow$ 1. Let $\tilde{\mathcal{M}}$ be the dense subspace of $\mathcal{H}$ formed from all finite linear combinations of $u_1, u_2, \cdots$. Then given $v \in \mathcal{H}$ and $\epsilon > 0$ there exists a $\sum_{n=1}^{N} \alpha_n u_n \in \tilde{\mathcal{M}}$ such that $||v - \sum_{n=1}^{N} \alpha_n u_n|| < \epsilon$. Q.E.D.

2.6 Linear operators and matrices, Least squares approximations

Let V, W be vector spaces over F (either the real or the complex field).

Definition 21 *A linear transformation (or linear operator) from V to W is a function $\mathbf{T} : V \to W$, defined for all $v \in V$ that satisfies $\mathbf{T}(\alpha u + \beta v) = \alpha \mathbf{T} u + \beta \mathbf{T} v$ for all $u, v \in V$, $\alpha, \beta \in F$. Here, the set $R(\mathbf{T}) = \{\mathbf{T} u : u \in V\}$ is called the range of $\mathbf{T}$.*

Lemma 8 *$R(\mathbf{T})$ is a subspace of W.*

PROOF: Let $w = \mathbf{T}u, z = \mathbf{T}v \in R(\mathbf{T})$ and let $\alpha, \beta \in F$. Then $\alpha w + \beta z = \mathbf{T}(\alpha u + \beta v) \in R(\mathbf{T})$. Q.E.D.

If V is m-dimensional with basis $v_1, \cdots, v_m$ and W is n-dimensional with basis $w_1, \cdots, w_n$ then $\mathbf{T}$ is completely determined by its matrix representation $T = (T_{jk})$ with respect to these two bases:

$$\mathbf{T} v_k = \sum_{j=1}^{n} T_{jk} w_j, \qquad , k = 1, 2, \cdots, m.$$

If $v \in V$ and $v = \sum_{k=1}^{m} \alpha_k v_k$ then the action $\mathbf{T} v = w$ is given by

$$\mathbf{T} v = \mathbf{T}(\sum_{k=1}^{m} \alpha_k v_k) = \sum_{k=1}^{m} \alpha_k \mathbf{T} v_k = \sum_{j=1}^{n} \sum_{k=1}^{m} (T_{jk} \alpha_k) w_j = \sum_{j=1}^{n} \beta_j w_j = w$$

Thus the coefficients β_j of w are given by $\beta_j = \sum_{k=1}^{m} T_{jk}\alpha_k$, $j = 1, \cdots, n$. In matrix notation, one writes this as

$$\begin{pmatrix} T_{11} & \cdots & T_{1m} \\ \vdots & \ddots & \vdots \\ T_{n1} & \cdots & T_{nm} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix},$$

or

$$Ta = b.$$

The matrix $T = (T_{jk})$ has n rows and m columns, i.e., it is $n \times m$, whereas the vector $a = (\alpha_k)$ is $m \times 1$ and the vector $b = (\beta_j)$ is $n \times 1$. If V and W are Hilbert spaces with ON bases, we shall sometimes represent operators $\mathbf{T}$ by matrices with an infinite number of rows and columns.

Let V, W, X be vector spaces over F, and $\mathbf{T}$, $\mathbf{U}$ be linear operators $\mathbf{T} : V \to W$,$\mathbf{U} : W \to X$. The *product* $\mathbf{UT}$ of these two operators is the composition $\mathbf{U} : V \to X$ defined by $\mathbf{UT}v = \mathbf{U}(\mathbf{T}v)$ for all $v \in V$.

Suppose V is m-dimensional with basis $v_1, \cdots, v_m$, W is n-dimensional with basis $w_1, \cdots, w_n$ and X is p-dimensional with basis $x_1, \cdots, x_p$. Then $\mathbf{T}$ has matrix representation $T = (T_{jk})$, $\mathbf{U}$ has matrix representation $U = (U_{\ell j})$,

$$\mathbf{U}w_j = \sum_{\ell=1}^{p} U_{\ell j} x_\ell, \qquad , j = 1, 2, \cdots, n,$$

and $\mathbf{Y} = \mathbf{UT}$ has matrix representation $Y = (Y_{\ell k})$ given by

$$\mathbf{Y}v_k = \mathbf{UT}v_k = \sum_{\ell=1}^{p} Y_{\ell k} x_\ell, \qquad k = 1, 2, \cdots, m,$$

A straightforward computation gives $Y_{\ell k} = \sum_{j=1}^{n} U_{\ell j} T_{jk}$, $\ell = 1, \cdots, p$, $k = 1, \cdots, m$. In matrix notation, one writes this as

$$\begin{pmatrix} U_{11} & \cdots & U_{1n} \\ \vdots & \ddots & \vdots \\ U_{p1} & \cdots & U_{pn} \end{pmatrix} \begin{pmatrix} T_{11} & \cdots & T_{1m} \\ \vdots & \ddots & \vdots \\ T_{n1} & \cdots & T_{nm} \end{pmatrix} = \begin{pmatrix} Y_{11} & \cdots & Y_{1m} \\ \vdots & \ddots & \vdots \\ Y_{p1} & \cdots & Y_{pm} \end{pmatrix},$$

or

$$UT = Y.$$

Here, U is $p \times n$, T is $n \times m$ and Y is $p \times m$.

Now let us return to our operator $\mathbf{T} : V \to W$ and suppose that both V and W are complex inner product spaces, with inner products $(\cdot, \cdot)_V, (\cdot, \cdot)_W$, respectively. Then $\mathbf{T}$ induces a linear operator $\mathbf{T}^* : W \to V$ and defined by

$$(\mathbf{T}v, w)_W = (v, \mathbf{T}^* w)_V, \qquad v \in V, w \in W.$$

To show that $\mathbf{T}^*$ exists, we will compute its matrix T^*. Suppose that $v_1, \cdots, v_m$ is an ON basis for V and $w_1, \cdots, w_n$ is an ON basis for W. Then we have have

$$T_{jk} = (\mathbf{T}v_k, w_j)_W = (v_k, \mathbf{T}^* w_j)_V = \bar{T}^*{}_{kj}, \quad k = 1, \cdots, m, \quad j = 1, \cdots, n.$$

Thus the operator $\mathbf{T}^*$, (the *adjoint operator* to $\mathbf{T}$) has the adjoint matrix to T: $T^*_{kj} = \bar{T}_{jk}$. In matrix notation this is written $T^* = \bar{T}^{\rm tr}$ where the ${}^{\rm tr}$ stands for the matrix transpose (interchange of rows and columns). For a real inner product space the complex conjugate is dropped and the adjoint matrix is just the transpose.

There are some special operators and matrices that we will meet often in this course. Suppose that $v_1, \cdots, v_m$ is an ON basis for V. The *identity operator* $\mathbf{I} : V \to V$ is defined by $\mathbf{I}v = v$ for all $v \in V$. The matrix of $\mathbf{I}$ is $I = (\delta_{jh})$ where $\delta_{jj} = 1$ and $\delta_{jh} = 0$ if $j \neq h$, $1 \leq j, h \leq m$. The *zero operator* $\mathbf{Z} : V \to V$ is defined by $\mathbf{Z}v = \Theta$ for all $v \in V$. The $n \times n$ matrix of $\mathbf{Z}$ has all matrix elements 0. An operator $\mathbf{U} : V \to V$ that preserves the inner product, $(\mathbf{U}v, \mathbf{U}u) = (v, u)$ for all $u, v \in V$ is called *unitary*. The matrix U of a unitary operator is characterized by the matrix equation $UU^* = I$. If V is a real inner product space, the operators $\mathbf{O} : V \to V$ that preserve the inner product, $(\mathbf{O}v, \mathbf{O}u) = (v, u)$ for all $u, v \in V$ are called *orthogonal*. The matrix O of an orthogonal operator is characterized by the matrix equation $OO^{tr} = I$.

2.6.1 Bounded operators on Hilbert spaces

In this section we present a few concepts and results from functional analysis that are needed for the study of wavelets.

An operator $\mathbf{T} : \mathcal{H} \to \mathcal{K}$ of the Hilbert space $\mathcal{H}$ to the Hilbert Space $\mathcal{K}$ is said to be *bounded* if it maps the unit ball $||u||_{\mathcal{H}} \leq 1$ to a bounded set in $\mathcal{K}$. This means that there is a finite positive number N such that

$$||\mathbf{T}u||_{\mathcal{K}} \leq N \quad \text{whenever} \quad ||u||_{\mathcal{H}} \leq 1.$$

The *norm* $||\mathbf{T}||$ of a bounded operator is its least bound:

$$||\mathbf{T}|| = \sup_{||u||_{\mathcal{H}} \leq 1} ||\mathbf{T}u||_{\mathcal{K}} = \sup_{||u||_{\mathcal{H}} = 1} ||\mathbf{T}u||_{\mathcal{K}}. \tag{2.4}$$

Lemma 9 *Let $\mathbf{T} : \mathcal{H} \to \mathcal{K}$ be a bounded operator.*

1. *$||\mathbf{T}u||_{\mathcal{K}} \leq ||\mathbf{T}|| \cdot ||u||_{\mathcal{H}}$ for all $u \in \mathcal{H}$.*

2. *If $\mathbf{S} : \mathcal{L} \to \mathcal{H}$ is a bounded operator from the Hilbert space $\mathcal{L}$ to $\mathcal{H}$, then $\mathbf{TS} : \mathcal{L} \to \mathcal{K}$ is a bounded operator with $||\mathbf{TS}|| \leq ||\mathbf{T}|| \cdot ||\mathbf{S}||$.*

PROOF:

1. The result is obvious for $u = \theta$. If u is nonzero, then $v = ||u||_{\mathcal{H}}^{-1}u$ has norm 1. Thus $||\mathbf{T}v||_{\mathcal{K}} \leq ||\mathbf{T}||$. The result follows from multiplying both sides of the inequality by $||u||_{\mathcal{H}}$.

2. ¿From part 1, $||\mathbf{TS}w||_{\mathcal{K}} = ||\mathbf{T}(\mathbf{S}w)||_{\mathcal{K}} \leq ||\mathbf{T}||\cdot||\mathbf{S}w||_{\mathcal{H}} \leq ||\mathbf{T}||\cdot||\mathbf{S}||\cdot||w||_{\mathcal{L}}$. Hence $||\mathbf{TS}|| \leq ||\mathbf{T}|| \cdot ||\mathbf{S}||$.

Q.E.D.

A special bounded operator is the *bounded linear functional* $\mathbf{f} : \mathcal{H} \to C$, where C is the one-dimensional vector space of complex numbers (with the absolute value
$cdot|$ as the norm). Thus $\mathbf{f}(u)$ is a complex number for each $u \in \mathcal{H}$ and $\mathbf{f}(\alpha u + \beta v) = \alpha\mathbf{f}(u) + \beta\mathbf{f}(v)$ for all scalars α, β and $u, v \in \mathcal{H}$ The *norm* of a bounded linear functional is defined in the usual way:

$$||\mathbf{f}|| = \sup_{||u||_{\mathcal{H}}=1} |\mathbf{f}(u). \tag{2.5}$$

For fixed $v \in \mathcal{H}$ the inner product $\mathbf{f}(u) \equiv (u, v)$, where $(\cdot, \cdot)$ is an import example of a bounded linear functional. The linearity is obvious and the functional is bounded since $|\mathbf{f}(u)| = |(u, v)| \leq ||u|| \cdot ||v||$. Indeed it is easy to show that $||\mathbf{f}|| = ||v||$. A very useful fact is that all bounded linear functionals on Hilbert spaces can be represented as inner products. This important result, the Riesz representation theorem, relies on the fact that a Hilbert space is complete. It is an elegant application of the projection theorem.

Theorem 17 *(Riesz representation theorem) Let $\mathbf{f}$ be a bounded linear functional on the Hilbert space $\mathcal{H}$. Then there is a vector $v \in \mathcal{H}$ such that $\mathbf{f}(u) = (u, v)$ for all $u \in \mathcal{H}$.*

PROOF:

- Let $\mathcal{N} = \{w \in \mathcal{H} : \mathbf{f}(w) = \theta\}$ be the null space of $\mathbf{f}$. Then $\mathcal{N}$ is a closed linear subspace of $\mathcal{H}$. Indeed if $w_1, w_2 \in \mathcal{N}$ and $\alpha, \beta \in C$ we have $\mathbf{f}(\alpha w_1 + \beta w_2) = \alpha\mathbf{f}(w_1) + \beta\mathbf{f}(w_2) = \theta$, so $\alpha w_1 + \beta w_2 \in \mathcal{N}$. If $\{w_n\}$ is a Cauchy sequence of vectors in $\mathcal{N}$, i.e., $\mathbf{f}(w_n) = \theta$, with $w_n \to w \in \mathcal{H}$ as $n \to \infty$ then

$$|\mathbf{f}(w)| = |\mathbf{f}(w) - \mathbf{f}(w_n)| = |\mathbf{f}(w - w_n)| \leq ||\mathbf{f}|| \cdot ||w - w_n|| \to 0$$

as $n \to \infty$. Thus $\mathbf{f}(w) = \theta$ and $w \in \mathcal{N}$, so $\mathcal{N}$ is closed.

- If $\mathbf{f}$ is the zero functional, then the theorem holds with $v = \theta$, the zero vector. If $\mathbf{F}$ is not zero, then there is a vector $u_0 \in \mathcal{H}$ such that $\mathbf{f}(u_0) = 1$. By the projection theorem we can decompose u_0 uniquely in the form $u_0 = v_0 + w_0$ where $w_0 \in \mathcal{N}$ and $v_0 \perp \mathcal{N}$. Then $1 = \mathbf{f}(u_0) = \mathbf{f}(v_0) + \mathbf{f}(w_0) = \mathbf{f}(v_0)$.

- Every $u \in \mathcal{H}$ can be expressed uniquely in the form $u = \mathbf{f}(u)v_0 + w$ for $w \in \mathcal{N}$. Indeed $\mathbf{f}(u - \mathbf{f}(u)v_0) = \mathbf{f}(u) - \mathbf{f}(u)\mathbf{f}(v_0) = 0$ so $u - \mathbf{f}(u)v_0 \in \mathcal{N}$.

- Let $v = ||v_0||^{-2}v_0$. Then $v \perp \mathcal{N}$ and

$$(u, v) = (\mathbf{f}(u)v_0 + w, v) = \mathbf{f}(u)(v_0, v) = \mathbf{f}(u)||v_0||^{-2}(v_0, v_0) = \mathbf{f}(u).$$

Q.E.D.

We can define adjoints of bounded operators on general Hilbert spaces, in analogy with our construction of adjoints of operators on finite-dimensional inner product spaces. We return to our bounded operator $\mathbf{T} : \mathcal{H} \to \mathcal{K}$. For any $v \in \mathcal{K}$ we define the linear functional $\mathbf{f}_v(u) = (\mathbf{T}u, v)_{\mathcal{K}}$ on $\mathcal{H}$. The functional is bounded because for $||u||_{\mathcal{H}} = 1$ we have

$$|\mathbf{f}_v(u)| = |(\mathbf{T}u, v)_{\mathcal{K}}| \leq ||\mathbf{T}u||_{\mathcal{K}} \cdot ||v||_{\mathcal{K}} \leq ||\mathbf{T}|| \cdot ||v||_{\mathcal{K}}.$$

By theorem 17 there is a unique vector $v^* \in \mathcal{H}$ such that

$$\mathbf{f}_v(u) \equiv (\mathbf{T}u, v)_{\mathcal{K}} = (u, v^*)_{\mathcal{H}},$$

for all $u \in \mathcal{H}$. We write this element as $v^* = \mathbf{T}^*v$. Thus $\mathbf{T}$ induces an operator $\mathbf{T}^* : \mathcal{K} \to \mathcal{H}$ and defined uniquely by

$$(\mathbf{T}u, v)_{\mathcal{K}} = (u, \mathbf{T}^*v)_{\mathcal{H}}, \qquad v \in \mathcal{H}, w \in \mathcal{K}.$$

Lemma 10 *1.* $\mathbf{T}^*$ *is a linear operator from* $\mathcal{K}$ *to* $\mathcal{H}$*.*

2. $\mathbf{T}^*$ *is a bounded operator.*

3. $||\mathbf{T}^*||^2 = ||\mathbf{T}||^2 = ||\mathbf{T}\mathbf{T}^*|| = ||\mathbf{T}^*\mathbf{T}||$.

PROOF:

1. Let $v \in \mathcal{K}$ and $\alpha \in C$. Then

$$(u, \mathbf{T}^*\alpha v)_{\mathcal{H}} = (\mathbf{T}u, \alpha v)_{\mathcal{K}} = \overline{\alpha}(\mathbf{T}u, v)_{\mathcal{K}} = \overline{\alpha}(u, \mathbf{T}^*v)_{\mathcal{H}}$$

so $\mathbf{T}^*(\alpha v) = \alpha\mathbf{T}^*v$. Now let $v_1, v_2 \in \mathcal{K}$. Then

$(u, \mathbf{T}^*[v_1+v_2])_{\mathcal{H}} = (\mathbf{T}u, [v_1+v-2])_{\mathcal{K}} = (\mathbf{T}u, v_1)_{\mathcal{K}}+(\mathbf{T}u, v_2)_{\mathcal{K}} = (u, \mathbf{T}^*v_+\mathbf{T}^*v_2)_{\mathcal{H}}$

so $\mathbf{T}^*(v_1 + v_2) = \mathbf{T}^*v_1 + \mathbf{T}^*v_2$.

2. Set $u = \mathbf{T}^*v$ in the defining equation $(\mathbf{T}u, v)_{\mathcal{K}} = (u, \mathbf{T}^*v)_{\mathcal{H}}$. Then

$||\mathbf{T}^*v||^2_{\mathcal{H}} = (\mathbf{T}^*v, \mathbf{T}^*v)_{\mathcal{H}} = (\mathbf{T}\mathbf{T}^*v, v)_{\mathcal{K}} \leq ||\mathbf{T}\mathbf{T}^*v||_{\mathcal{K}}||v||_{\mathcal{K}} \leq ||\mathbf{T}||\cdot||\mathbf{T}^*v||_{\mathcal{H}}||v||_{\mathcal{K}}.$

Canceling the common factor $||\mathbf{T}^*v||_{\mathcal{H}}$ from the far left and far right-hand sides of these inequalities, we obtain

$$||\mathbf{T}^*v||_{\mathcal{H}} \leq ||\mathbf{T}|| \cdot ||v||_{\mathcal{K}},$$

so $\mathbf{T}^*$ is bounded.

3. From the last inequality of the proof of 2 we have $||\mathbf{T}^*|| \leq ||T||$. However, if we set $v = \mathbf{T}u$ in the defining equation $(\mathbf{T}u, v)_{\mathcal{K}} = (u, \mathbf{T}^*v)_{\mathcal{H}}$, then we obtain an analogous inequality

$$||\mathbf{T}u||_{\mathcal{K}} \leq ||\mathbf{T}^*|| \cdot ||u||_{\mathcal{H}}.$$

This implies $||\mathbf{T}|| \leq ||bfT^*||$. Thus $||\mathbf{T}|| = ||bfT^*||$. From the proof of part 2 we have

$$||\mathbf{T}^*v||^2_{\mathcal{H}} = (\mathbf{T}\mathbf{T}^*v, v)_{\mathcal{K}}. \qquad (2.6)$$

Applying the Schwarz inequalty to the right-hand side of this identity we have

$$||\mathbf{T}^*v||^2_{\mathcal{H}} \leq ||\mathbf{T}\mathbf{T}^*v||_{\mathcal{K}}||v||_{\mathcal{K}} \leq ||\mathbf{T}\mathbf{T}^*|| \cdot ||v||^2_{\mathcal{K}},$$

so $||\mathbf{T}^*||^2 \leq ||\mathbf{T}\mathbf{T}^*||$. But from lemma 9 we have $||\mathbf{T}\mathbf{T}^*|| \leq ||\mathbf{T}|| \cdot ||\mathbf{T}^*||$, so

$$||\mathbf{T}^*||^2 \leq ||\mathbf{T}\mathbf{T}^*|| \leq ||\mathbf{T}|| \cdot ||\mathbf{T}^*||| = ||\mathbf{T}^*||^2.$$

An analogous proof, switching the roles of u and v, yields

$$||\mathbf{T}||^2 \leq ||\mathbf{T}^*\mathbf{T}|| \leq ||\mathbf{T}|| \cdot ||\mathbf{T}^*||| = ||\mathbf{T}||^2.$$

Q.E.D.

2.6.2 Least squares approximations

Many applications of mathematics to statistics, image processing, numerical analysis, global positioning systems, etc., reduce ultimately to solving a system of equations of the form $Ta = b$ or

$$\begin{pmatrix} T_{11} & \cdots & T_{1m} \\ \vdots & \ddots & \vdots \\ T_{n1} & \cdots & T_{nm} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \qquad (2.7)$$

Here $b = \{\beta_1, \cdots, \beta_n\}$ are n measured quantities, the $n \times m$ matrix $T = (T_{jk})$ is known, and we have to compute the m quantities $a = \{\alpha_1, \cdots, \alpha_m\}$. Since b is measured experimentally, there may be errors in these quantities. This will induce errors in the calculated vector a. Indeed for some measured values of b there may be no solution a.

EXAMPLE: Consider the 3 × 2 system

$$\begin{pmatrix} 3 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ \beta_3 \end{pmatrix}.$$

If $\beta_3 = 5$ then this system has the unique solution $\alpha_1 = -1, \alpha_2 = 3$. However, if $\beta_3 = 5 + \epsilon$ for ϵ small but nonzero, then there is no solution!

We want to guarantee an (approximate) solution of (2.7) for all vectors b and matrices T. We adopt a least squares approach. Let's embed our problem into the inner product spaces V and W above. That is T is the matrix of the operator $\mathbf{T} : V \to W$, b is the component vector of a given $w \in W$ (with respect to the $\{w_j\}$ basis), and a is the component vector of $v \in V$ (with respect to the $\{v_k\}$

basis), which is to be computed. Now the original equation $Ta = b$ becomes $\mathbf{T}v = w$.

Let us try to find an approximate solution v of the equation $\mathbf{T}v = w$ such that the norm of the error $||w - \mathbf{T}v||_W$ is minimized. If the original problem has an exact solution then the error will be zero; otherwise we will find a solution v_0 with minimum (least squares) error. The square of the error will be

$$\epsilon^2 = \min_{v \in V} ||w - \mathbf{T}v||^2_W = ||w - \mathbf{T}v_0||^2_W$$

This may not determine v_0 uniquely, but it will uniquely determine $\mathbf{T}v_0$.

We can easily solve this problem via the projection theorem. recall that the range of $\mathbf{T}$, $R(\mathbf{T}) = \{\mathbf{T}u : u \in V\}$ is a subspace of W. We need to find the point on $R(\mathbf{T})$ that is closest in norm to w. By the projection theorem, that point is just the projection of w on $R(\mathbf{T})$, i.e., the point $\mathbf{T}v_0 \in R(\mathbf{T})$ such that $w - \mathbf{T}v_0 \perp R(\mathbf{T})$. This means that

$$(w - \mathbf{T}v_0, \mathbf{T}v)_W = 0$$

for all $v \in V$. Now, using the adjoint operator, we have

$$(w - \mathbf{T}v_0, \mathbf{T}v)_W = (\mathbf{T}^*[w - \mathbf{T}v_0], v)_V = (\mathbf{T}^*w - \mathbf{T}^*\mathbf{T}v_0, v)_V = 0$$

for all $v \in V$. This is possible if and only if

$$\mathbf{T}^*\mathbf{T}v_0 = \mathbf{T}^*w.$$

In matrix notation, our equation for the least squares solution a_0 is

$$T^*Ta_0 = T^*b. \tag{2.8}$$

The original system was rectangular; it involved m equations for n unknowns. Furthermore, in general it had no solution. Here however, the $n \times n$ matrix T^*T is square and the are n equations for the n unknowns $a_0 = \{\alpha_1, \cdots, \alpha_n\}$. If the matrix T is real, then equations (2.8) become $T^{\rm tr}Ta_0 = T^{\rm tr}b$. This problem ALWAYS has a solution a_0 and Ta_0 is unique.

There is a nice example of the use of the least squares approximation in linear predictive coding (see Chapter 0 of Boggess and Narcowich). This is a data compression algorithm used to eliminate partial redundancy in a signal. We will revisit the least squares approximation in the study of Fourier series and wavelets.

Chapter 3

Fourier Series

3.1 Definitions, Real and complex Fourier series

We have observed that the functions $e_n(t) = e^{int}/\sqrt{2\pi}$, $n = 0, \pm 1, \pm 2, \cdots$ form an ON set for in the Hilbert space $L^2[0, 2\pi]$ of square-integrable functions on the interval $[0, 2\pi]$. In fact we shall show that these functions form an ON basis. Here the inner product is

$$(u, v) = \int_0^{2\pi} u(t)\overline{v}(t)\, dt, \qquad u, v \in L^2[0, 2\pi].$$

We will study this ON set and the completeness and convergence of expansions in the basis, both pointwise and in the norm. Before we get started, it is convenient to assume that $L^2[0, 2\pi]$ consists of square-integrable functions on the unit circle, rather than on an interval of the real line. Thus we will replace every function $f(t)$ on the interval $[0, 2\pi]$ by a function $f^*(t)$ such that $f^*(0) = f^*(2\pi)$ and $f^*(t) = f(t)$ for $0 \leq t < 2\pi$. Then we will extend f^* to all $-\infty < t < \infty$ be requiring periodicity: $f^*(t + 2\pi) = f^*(t)$. This will not affect the values of any integrals over the interval $[0, 2\pi]$. Thus, from now on our functions will be assumed $2\pi - periodic$. One reason for this assumption is the

Lemma 11 *Suppose F is $2\pi - periodic$ and integrable. Then for any real number a*

$$\int_a^{2\pi+a} F(t)dt = \int_o^{2\pi} F(t)dt.$$

NOTE: Each side of the identity is just the integral of F over one period. For an analytic proof we use the Fundamental Theorem of Calculus and the chain rule:

$$\frac{d}{da}\int_a^{2\pi+a} F(t)dt = F(t)|_a^{2\pi+a}\frac{da}{da} = F(2\pi+a) - F(a) = 0,$$

so $\int_a^{2\pi+a} F(t)dt$ is a constant independent of a.

Thus we can transfer all our integrals to any interval of length 2π without altering the results.

For students who don't have a background in complex variable theory we will define the complex exponential in terms of real sines and cosines, and derive some of its basic properties directly. Let $z = x + iy$ be a complex number, where x and y are real. (Here and in all that follows, $i = \sqrt{-1}$.) Then $\bar{z} = x - iy$.

Definition 22 $e^z = \exp(x)(\cos y + i \sin y)$

Lemma 12 *Properties of the complex exponential:*

- $e^{z_1}e^{z_2} = e^{z_1+z_2}$
- $|e^z| = \exp(x)$
- $\overline{e^z} = e^{\overline{z}} = \exp(x)(\cos y - i \sin y)$.

Simple consequences for the basis functions $e_n(t) = e^{int}/\sqrt{2\pi},\ n = 0, \pm1, \pm2, \cdots$ where t is real, are given by

Lemma 13 *Properties of e^{int}:*

- $e^{in(t+2\pi)} = e^{int}$
- $|e^{int}| = 1$
- $\overline{e^{int}} = e^{-int}$
- $e^{imt}e^{int} = e^{i(m+n)t}$
- $e^0 = 1$
- $\frac{d}{dt}e^{int} = ine^{int}$.

Lemma 14 $(e_n, e_m) = \delta_{nm}$.

PROOF: If $n \neq m$ then

$$(e_n, e_m) = \frac{1}{2\pi}\int_0^{2\pi} e^{i(n-m)t}dt = \frac{1}{2\pi}\frac{e^{i(n-m)t}}{i(n-m)}\,|_0^{2\pi} = 0.$$

If $n = m$ then $(e_n, e_m) = \frac{1}{2\pi}\int_0^{2\pi} 1\, dt = 1$. Q.E.D.

Since $\{e_n\}$ is an ON set, we can project any $f \in L^2[0, 2\pi]$ on the subspace generated by this set to get the Fourier expansion

$$f(t) \sim \sum_{n=-\infty}^{\infty} (f, e_n)e_n(t),$$

or

$$f(t) \sim \sum_{n=-\infty}^{\infty} c_n e^{int}, \qquad c_n = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-int}dt. \tag{3.1}$$

This is the *complex version* of Fourier series. (For now the $\sim$ just denotes that the right-hand side is the Fourier series of the left-hand side. In what sense the Fourier series represents the function is a matter to be resolved.) From our study of Hilbert spaces we already know that Bessel's inequality holds: $(f, f) \geq \sum_{n=-\infty}^{\infty} |(f, e_n)|^2$ or

$$\frac{1}{2\pi}\int_0^{2\pi} |f(t)|^2 dt \geq \sum_{n=-\infty}^{\infty} |c_n|^2. \tag{3.2}$$

An immediate consequence is the Riemann-Lebesgue Lemma.

Lemma 15 *(Riemann-Lebesgue, weak form)* $\lim_{|n|\to\infty} \int_0^{2\pi} f(t)e^{-int}dt = 0.$

Thus, as $|n|$ gets large the Fourier coefficients go to 0.

If f is a real-valued function then $\bar{c}_n = c_{-n}$ for all n. If we set

$$c_n = \frac{a_n - ib_n}{2}, \qquad n = 0, 1, 2, \cdots$$

$$c_{-n} = \frac{a_n + ib_n}{2}, \qquad n = 1, 2, \cdots$$

and rearrange terms, we get the *real version* of Fourier series:

$$\begin{aligned} f(t) &\sim \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt), \qquad a_n = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos nt\, dt \\ b_n &= \frac{1}{\pi}\int_0^{2\pi} f(t)\sin nt\, dt \end{aligned} \tag{3.3}$$

with Bessel inequality

$$\frac{1}{\pi}\int_0^{2\pi}|f(t)|^2dt \geq \frac{|a_0|^2}{2}+\sum_{n=1}^{\infty}(|a_n|^2+|b_n|^2).$$

REMARK: The set $\{\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos nt, \frac{1}{\sqrt{\pi}}\sin nt\}$ for $n = 1, 2, \cdots$ is also ON in $L^2[0, 2\pi]$, as is easy to check, so (3.3) is the correct Fourier expansion in this basis for complex functions $f(t)$, as well as real functions.

Later we will prove the following basic results:

Theorem 18 *Parseval's equality. Let $f \in L^2[0, 2\pi]$. Then $(f, f) = \sum_{n=-\infty}^{\infty}|(f, e_n)|^2$.*

In terms of the complex and real versions of Fourier series this reads

$$\frac{1}{2\pi}\int_0^{2\pi}|f(t)|^2dt = \sum_{n=-\infty}^{\infty}|c_n|^2 \tag{3.4}$$

or

$$\frac{1}{\pi}\int_0^{2\pi}|f(t)|^2dt = \frac{|a_0|^2}{2}+\sum_{n=1}^{\infty}(|a_n|^2+|b_n|^2).$$

Let $f \in L^2[0, 2\pi]$ and remember that we are assuming that all such functions satisfy $f(t+2\pi) = f(t)$. We say that f is *piecewise continuous* on $[0, 2\pi]$ if it is continuous except for a finite number of discontinuities. Furthermore, at each t the limits $f(t+0) = \lim_{h\to 0, h>0} f(t+h)$ and $f(t-0) = \lim_{h\to 0, h>0} f(t-h)$ exist. NOTE: At a point t of continuity of f we have $f(t+0) = f(t-0)$, whereas at a point of discontinuity $f(t+0) \neq f(t-0)$ and $f(t+0) - f(t-0)$ is the magnitude of the jump discontinuity.

Theorem 19 *Suppose*

- *$f(t)$ is periodic with period 2π.*
- *$f(t)$ is piecewise continuous on $[0, 2\pi]$.*
- *$f'(t)$ is piecewise continuous on $[0, 2\pi]$.*

Then the Fourier series of $f(t)$ converges to $\frac{f(t+0)+f(t-0)}{2}$ at each point t.

3.2 Examples

We will use the real version of Fourier series for these examples. The transformation to the complex version is elementary.

1. Let

$$f(t) = \begin{cases} 0, & t = 0 \\ \frac{\pi - t}{2}, & 0 < t < 2\pi \\ 0, & t = 2\pi. \end{cases}$$

and $f(t + 2\pi) = f(t)$. We have $a_0 = \frac{1}{\pi} \int_0^{2\pi} \frac{\pi - t}{2} \, dt = 0$. and for $n \geq 1$,

$$a_n = \frac{1}{\pi} \int_0^{2\pi} \frac{\pi - t}{2} \cos nt \, dt = \frac{\frac{\pi - t}{2} \sin nt}{n\pi} \Big|_0^{2\pi} + \frac{1}{2\pi n} \int_0^{2\pi} \sin nt \, dt = 0,$$

$$b_n = \frac{1}{\pi} \int_0^{2\pi} \frac{\pi - t}{2} \sin nt \, dt = -\frac{\frac{\pi - t}{2} \cos nt}{n\pi} \Big|_0^{2\pi} - \frac{1}{2\pi n} \int_0^{2\pi} \cos nt \, dt = \frac{1}{n}.$$

Therefore,

$$\frac{\pi - t}{2} = \sum_{n=1}^{\infty} \frac{\sin nt}{n}, \qquad 0 < t < 2\pi.$$

By setting $t = \pi/2$ in this expansion we get an alternating series for $\pi/4$:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots.$$

Parseval's identity gives

$$\frac{\pi^2}{6} = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

2. Let

$$f(t) = \begin{cases} \frac{1}{2}, & t = 0 \\ 1, & 0 < t < \pi \\ \frac{1}{2}, & t = \pi \\ 0 & \pi < t < 2\pi. \end{cases}$$

and $f(t + 2\pi) = f(t)$ (a step function). We have $a_0 = \frac{1}{\pi} \int_0^{\pi} dt = 1$, and for $n \geq 1$,

$$a_n = \frac{1}{\pi} \int_0^{\pi} \cos nt \, dt = \frac{\sin nt}{n\pi} \Big|_0^{\pi} = 0,$$

$$b_n = \frac{1}{\pi} \int_0^{\pi} \sin nt \, dt = -\frac{\cos nt}{n\pi} \Big|_0^{\pi} = \frac{(-1)^{n+1} + 1}{n\pi} = \begin{cases} \frac{2}{\pi n}, & n \text{ odd} \\ 0, & n \text{ even.} \end{cases}$$

Therefore,

$$f(t) = \frac{1}{2} + \frac{2}{\pi}\sum_{j=1}^{\infty} \frac{\sin(2j-1)t}{2j-1}.$$

For $0 < t < \pi$ this gives

$$\frac{\pi}{4} = \sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + \cdots,$$

and for $\pi < t < 2\pi$ it gives

$$-\frac{\pi}{4} = \sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + \cdots.$$

Parseval's equality becomes

$$\frac{\pi^2}{8} = \sum_{j=1}^{\infty} \frac{1}{(2j-1)^2}.$$

3.3 Fourier series on intervals of varying length, Fourier series for odd and even functions

Although it is convenient to base Fourier series on an interval of length 2π there is no necessity to do so. Suppose we wish to look at functions $f(x)$ in $L^2[\alpha,\beta]$. We simply make the change of variables $t = \frac{2\pi x}{\beta-\alpha}$ in our previous formulas. Every function $f(x) \in L^2[\alpha,\beta]$ is uniquely associated with a function $\hat{f}(t) \in L^2[0,2\pi]$ by the formula $f(x) = \hat{f}(\frac{2\pi x}{\beta-\alpha})$. The set $\{\frac{1}{\sqrt{\beta-\alpha}}, \sqrt{\frac{2}{\beta-\alpha}}\cos\frac{2\pi nx}{\beta-\alpha}, \sqrt{\frac{2}{\beta-\alpha}}\sin\frac{2\pi nx}{\beta-\alpha}\}$ for $n = 1, 2, \cdots$ is an ON basis for $L^2[\alpha,\beta]$, The real Fourier expansion is

$$\begin{aligned} f(x) &\sim \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos\frac{2\pi nx}{\beta-\alpha} + b_n \sin\frac{2\pi nx}{\beta-\alpha}), \qquad (3.5) \\ a_n &= \frac{2}{\beta-\alpha}\int_{\alpha}^{\beta} f(x)\cos\frac{2\pi nx}{\beta-\alpha}\,dx, \quad b_n = \frac{2}{\beta-\alpha}\int_{\alpha}^{\beta} f(x)\sin\frac{2\pi nx}{\beta-\alpha}\,dx \end{aligned}$$

with Parseval equality

$$\frac{2}{\beta-\alpha}\int_{\alpha}^{\beta} |f(x)|^2 dx = \frac{|a_0|^2}{2} + \sum_{n=1}^{\infty}(|a_n|^2 + |b_n|^2).$$

For our next variant of Fourier series it is convenient to consider the interval $[-\pi, \pi]$ and the Hilbert space $L^2[-\pi, \pi]$. This makes no difference in the formulas, since all elements of the space are 2π-periodic. Now suppose $f(t)$ is defined and square integrable on the interval $[0, \pi]$. We define $F(t) \in L^2[-\pi, \pi]$ by

$$F(t) = \begin{cases} f(t) & \text{on } [0, \pi] \\ f(-t) & \text{for } -\pi < t < 0 \end{cases}$$

The function F has been constructed so that it is *even*, i.e., $F(-t) = F(t)$. For an even functions the coefficients $b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} F(t) \sin nt\, dt = 0$ so

$$F(t) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt$$

on $[-\pi, \pi]$ or

$$f(t) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt, \quad \text{for } o \le t \le \pi \tag{3.6}$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} F(t) \cos nt\, dt = \frac{2}{\pi}\int_0^{\pi} f(t) \cos nt\, dt.$$

Here, (3.6) is called the *Fourier cosine series* of f.

We can also extend the function $f(t)$ from the interval $[0, \pi]$ to an odd function on the interval $[-\pi, \pi]$. We define $G(t) \in L^2[-\pi, \pi]$ by

$$G(t) = \begin{cases} f(t) & \text{on } (0, \pi] \\ 0 & \text{for } t = 0 \\ -f(-t) & \text{for } -\pi < t < 0. \end{cases}$$

The function G has been constructed so that it is *odd*, i.e., $G(-t) = -G(t)$. For an odd function the coefficients $a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} G(t) \cos nt\, dt = 0$ so

$$G(t) \sim \sum_{n=1}^{\infty} b_n \sin nt$$

on $[-\pi, \pi]$ or

$$f(t) \sim \sum_{n=1}^{\infty} b_n \sin nt, \quad \text{for } 0 < t \le \pi, \tag{3.7}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} G(t) \sin nt\, dt = \frac{2}{\pi}\int_0^{\pi} f(t) \sin nt\, dt.$$

Here, (3.7) is called the *Fourier sine series* of f.

EXAMPLE: $f(t) = t, \quad 0 \leq t \leq \pi$. Fourier Sine series.

$$b_n = \frac{2}{\pi}\int_0^\pi t \sin nt \, dt = \frac{-2t\cos nt}{n\pi}\,|_0^\pi + \frac{2}{n\pi}\int_0^\pi \cos nt \, dt = \frac{2(-1)^{n+1}}{n}.$$

Therefore,

$$t = \sum_{n=1}^\infty \frac{2(-1)^{n+1}}{n}\sin nt, \qquad 0 < t < \pi.$$

Fourier Cosine series.

$$a_n = \frac{2}{\pi}\int_0^\pi t \cos nt \, dt = \frac{2t\sin nt}{n\pi}\,|_0^\pi - \frac{2}{n\pi}\int_0^\pi \sin nt \, dt = \frac{2[(-1)^n - 1}{n^2\pi},$$

for $n \geq 1$ and $a_0 = \frac{2}{\pi}\int_0^\pi t \, dt = \pi$, so

$$t = \frac{\pi}{2} - \frac{4}{\pi}\sum_{j=1}^\infty \frac{\cos(2j-1)t}{(2j-1)^2}, \qquad 0 < t < \pi.$$

3.4 Convergence results

In this section we will prove the pointwise convergence Theorem 19. Let f be a complex valued function such that

- $f(t)$ is periodic with period 2π.
- $f(t)$ is piecewise continuous on $[0, 2\pi]$.
- $f'(t)$ is piecewise continuous on $[0, 2\pi]$.

Expanding f in a Fourier series (real form) we have

$$\begin{aligned} f(t) &\sim \frac{a_0}{2} + \sum_{n=1}^\infty (a_n \cos nt + b_n \sin nt) = S(t), \qquad a_n = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos nt \, dt \\ b_n &= \frac{1}{\pi}\int_0^{2\pi} f(t) \sin nt \, dt \end{aligned} \tag{3.8}$$

For a fixed t we want to understand the conditions under which the Fourier series converges to a number $S(t)$, and the relationship between this number and f. To be more precise, let

$$S_k(t) = \frac{a_0}{2} + \sum_{n=1}^{k} (a_n \cos nt + b_n \sin nt)$$

be the k-th partial sum of the Fourier series. This is a finite sum, a *trigonometric polynomial,* so it is well defined for all $t \in R$. Now we have

$$S(t) = \lim_{k \to \infty} S_k(t),$$

if the limit exists. To better understand the properties of $S_k(t)$ in the limit, we will recast this finite sum as a single integral. Substituting the expressions for the Fourier coefficients a_n, b_n into the finite sum we find

$$S_k(t) = \frac{1}{2\pi}\int_0^{2\pi} f(x)dx + \frac{1}{\pi}\sum_{n=1}^{k}\left(\int_0^{2\pi} f(x)\cos nx\, dx \cos nt + \int_0^{2\pi} f(x)\sin nx\, dx \sin nt\right),$$

so

$$\begin{aligned} S_k(t) &= \frac{1}{\pi}\int_0^{2\pi}\left[\frac{1}{2} + \sum_{n=1}^{k}(\cos nx \cos nt + \sin nx \sin nt\right] f(x)dx \\ &= \frac{1}{\pi}\int_0^{2\pi}\left[\frac{1}{2} + \sum_{n=1}^{k}\cos[n(t-x)]\right] f(x)dx \\ &= \frac{1}{\pi}\int_0^{2\pi} D_k(t-x)f(x)dx. \qquad (3.9) \end{aligned}$$

We can find a simpler form for the kernel $D_k(t) = \frac{1}{2} + \sum_{n=1}^{k}\cos nt = -\frac{1}{2} + \sum_{m=0}^{k}\cos mt$. The last cosine sum is the real part of the geometric series

$$\sum_{m=0}^{k}(e^{it})^m = \frac{(e^{it})^{k+1} - 1}{e^{it} - 1}$$

so

$$-\frac{1}{2} + \sum_{m=0}^{k}\cos mt = -\frac{1}{2} + \text{Re}\,\frac{(e^{it})^{k+1} - 1}{e^{it} - 1}$$

$$= \text{Re}\,\frac{(e^{it})^{k+1} - \frac{1}{2}e^{it} - \frac{1}{2}}{e^{it} - 1} = \text{Re}\,\frac{e^{ikt} - e^{i(k+1)t} + \frac{1}{2}(e^{it} - e^{-it})}{4\sin^2\frac{t}{2}}.$$

Thus,

$$D_k(t) = \frac{\cos kt - \cos(k+1)t}{4\sin^2 \frac{t}{2}} = \frac{\sin(k+\frac{1}{2})t}{2\sin\frac{t}{2}}. \tag{3.10}$$

Note that D has the properties:

- $D_k(t) = D_k(t+2\pi)$
- $D_k(-t) = D_k(t)$
- $D_k(t)$ is defined and differentiable for all t and $D_k(0) = k + \frac{1}{2}$.

¿From these properties it follows that the integrand of (3.9) is a 2π-periodic function of x, so that we can take the integral over any full 2π-period:

$$S_k(t) = \frac{1}{\pi}\int_{a-\pi}^{a+\pi} D_k(t-x)f(x)dx,$$

for any real number a. Let us set $a = t$ and fix a δ such that $0 < \delta < \pi$. (Think of δ as a very small positive number.) We break up the integral as follows:

$$S_k(t) = \frac{1}{\pi}\left(\int_{t-\pi}^{t-\delta} + \int_{t+\delta}^{t+\pi}\right) D_k(t-x)f(x)dx + \frac{1}{\pi}\int_{t-\delta}^{t+\delta} D_k(t-x)f(x)dx.$$

For fixed t we can write $D_k(t-x)$ in the form

$$\begin{aligned} D_k(t-x) &= \frac{f_1(x,t)\cos k(t-x) + f_2(x,t)\sin k(t-x)}{\sin[\frac{1}{2}(t-x)]} \\ &= g_1(x,t)\cos k(t-x) + g_2(x,t)\sin k(t-x). \end{aligned}$$

In the interval $[t-\pi, t-\delta]$ the functions g_1, g_2 are bounded. Thus the functions

$$G_\ell(x,t) = \begin{cases} g_\ell(x,t) & \text{for } x \in [t-\pi, t-\delta] \\ 0 & \text{elsewhere} \end{cases}, \qquad \ell = 1,2$$

are elements of $L^2[-\pi,\pi]$ (and its 2π-periodic extension). Thus, by the Riemann-Lebesgue Lemma, applied to the ON basis determined by the orthogonal functions $\cos k(t-x), \sin k(t-x)$, the first integral goes to 0 as $k \to \infty$. A similar argument shows that the integral over the interval $[t+\delta, t+\pi]$ goes to 0 as $k \to \infty$. [This argument doesn't hold for the interval $[t-\delta, t+\delta]$ because the term $\sin[\frac{1}{2}(t-x)]$ vanishes in the interval, so that the G_ℓ are not square integrable.] Thus,

$$\lim_{k\to\infty} S_k(t) = \lim_{k\to\infty}\frac{1}{\pi}\int_{t-\delta}^{t+\delta} D_k(t-x)f(x)dx, \tag{3.11}$$

where,

$$D_k(t) = \frac{\sin(k+\frac{1}{2})t}{2\sin\frac{t}{2}}.$$

Theorem 20 *Localization Theorem. The sum $S(t)$ of the Fourier series of f at t is completely determined by the behavior of f in an arbitrarily small interval $(t-\delta, t+\delta)$ about t.*

This is a remarkable fact! Although the Fourier coefficients contain information about all of the values of f over the interval $[0, 2\pi)$, only the local behavior of f affects the convergence at a specific point t.

3.4.1 The convergence proof: part 1

Using the properties of $D(t)$ derived above, we continue to manipulate the limit expression into a more tractable form:

$$\lim_{k\to\infty} S_k(t) = \lim_{k\to\infty}\frac{1}{\pi}\int_{t-\delta}^{t+\delta} D_k(t-x)f(x)dx = \lim_{k\to\infty}\frac{1}{\pi}\int_{-\delta}^{\delta} D_k(u)f(t+u)du$$

$$= \lim_{k\to\infty}\frac{1}{\pi}\int_0^{\delta} D_k(x)[f(t+x)+f(t-x)]dx.$$

Finally,

$$\begin{aligned}\lim_{k\to\infty} S_k(t) &= \lim_{k\to\infty}\frac{1}{\pi}\int_0^{\delta}\frac{\sin[(k+\frac{1}{2})x]}{x}\frac{[f(t+x)+f(t-x)]x}{2\sin\frac{x}{2}}dx \\ &= \lim_{k\to\infty}\frac{1}{\pi}\int_0^{\delta}\frac{\sin[(k+\frac{1}{2})x]}{x}F(x)dx \qquad (3.12)\end{aligned}$$

Here

$$F(x) = \frac{[f(t+x)+f(t-x)]x}{2\sin\frac{x}{2}}.$$

Properties of $F(x)$:

- $F(x)$ is piecewise continuous on $[0, \delta]$
- $F'(x)$ is piecewise continuous on $[0, \delta]$
- $F(+0) = f(t+) + f(t-)$

- $F'(+0) = f'(+0) - f'(-0)$ PROOF:

$$F'(+0) = \lim_{h\to 0, h>0} \frac{F(h) - F(+0)}{h}$$
$$= \lim_{h\to 0, h>0} \frac{f(t+h) + f(t-h) - f(t+0) - f(t-0)}{h}$$
$$= \lim_{h\to 0, h>0} \frac{f(t+h) - f(t+0)}{h} + \lim_{h\to 0, h>0} \frac{f(t-h) - f(t-0)}{h}$$
$$= f'(t+0) - f'(t-0).$$

Q.E.D.

Now we see what is going on! All the action takes place in a neighborhood of $x = 0$. The function $F(x)$ is well behaved near $x = 0$: $F(0+)$ and $F'(0+)$ exist. However the function $\frac{\sin[(k+\frac{1}{2})x]}{x}$ has a maximum value of $k + \frac{1}{2}$ at $x = 0$, which blows up as $k \to \infty$. Also, as k increases without bound, $\frac{\sin[(k+\frac{1}{2})x]}{x}$ decreases rapidly from its maximum value and oscillates more and more quickly.

3.4.2 Some important integrals

To finish up the convergence proof we need to do some separate calculations. First of all we will need the value of the improper Riemann integral $\int_0^\infty \frac{\sin x}{x} dx$. The function $\mathrm{sinc}\, x$ is one of the most important that occurs in this course, both in the theory and in the applications.

Definition 23 $\mathrm{sinc}\, x = \begin{cases} \frac{\sin \pi x}{\pi x} & \textit{for } x \neq 0 \\ 1 & \textit{for } x = 0. \end{cases}$

The sinc function is one of the few we will study that is *not* Lebesgue integrable. Indeed we will show that the $L^1[0,\infty]$ norm of the sinc function is infinite. (The $L^2[0,\infty]$ norm is finite.) However, the sinc function is improper Riemann integrable because it is related to a (conditionally) convergent alternating series. Computing the integral is easy if you know about contour integration. If not, here is a direct verification (with some tricks).

Lemma 16 *The sinc function doesn't belong to $L^1[0,\infty]$. However the improper Riemann integral $I = \lim_{N\to\infty} \int_0^{N\pi} \frac{\sin x}{x} dx$ does converge and*

$$I = \int_0^\infty \frac{\sin x}{x} dx = \pi \int_0^\infty \mathrm{sinc}\, x \, dx = \frac{\pi}{2}. \tag{3.13}$$

PROOF: Set $A_k = \int_{k\pi}^{(k+1)\pi} |\frac{\sin x}{x}|dx$. Note that

$$\frac{2}{(k+1)\pi} = \int_{k\pi}^{(k+1)\pi} \frac{|\sin x|}{(k+1)\pi}dx < \int_{k\pi}^{(k+1)\pi} |\frac{\sin x}{x}|dx < \int_{k\pi}^{(k+1)\pi} \frac{|\sin x|}{k\pi}dx| = \frac{2}{k\pi},$$

so $\frac{2}{(k+1)\pi} < A_k < \frac{2}{(k)\pi}$ for $k > 0$. Thus $A_k \to 0$ as $k \to \infty$ and $A_{k+1} < A_k$, so $I = \sum_{k=0}^{\infty}(-1)^k A_k$ converges. (Indeed, it is easy to see that the even sums $E_h = \sum_{k=0}^{2h}(-1)^k A_k$ form a decreasing sequence $E_0 > E_1 > \cdots > E_h > \cdots > I$ of upper bounds for I and the odd sums $O_h = \sum_{k=0}^{2h+1}(-1)^k A_k$ form an increasing sequence $O_0 < O_1 < \cdots < O_h < \cdots < I$ of lower bounds for I. Moreover $E_h - O_h = A_{2h+1} \to 0$ as $h \to \infty$.) However,

$$\int_{\pi}^{\infty} |\frac{\sin x}{x}|dx = \sum_{k=1}^{\infty} A_k > \frac{2}{\pi}(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k} + \cdots)$$

diverges, since the harmonic series diverges.

Now consider the function

$$G(t) = \int_0^{\infty} e^{-tx}\frac{\sin x}{x}dx, \tag{3.14}$$

defined for $\Re t \geq 0$. Note that

$$G(t) = \sum_{k=0}^{\infty}(-1)^k \int_{k\pi}^{(k+1)\pi} (-1)^k e^{-tx}\frac{\sin x}{x}dx$$

converges uniformly on $[0, \infty]$ which implies that $G(t)$ is continuous for $t \geq 0$ and infinitely differentiable for $t > 0$. Also $G(0) = \lim_{t\to 0, t>0} G(t) = G(+0)$.

Now, $G'(t) = -\int_0^{\infty} e^{-tx}\sin x\, dx$. Integrating by parts we have

$$\int_0^{\infty} e^{-tx}\sin x\, dx = \frac{-e^{-tx}\sin x}{t}|_0^{\infty} + \int_0^{\infty} e^{-tx}\cos x\, dx = \frac{-e^{-tx}\cos x}{t^2}|_0^{\infty}$$

$$-\int_0^{\infty} \frac{e^{-tx}\sin x}{t^2}dx.$$

Hence, $\int_0^{\infty} e^{-tx}\sin x\, dx = \frac{1}{1+t^2}$ and $G'(t) = \frac{-1}{1+t^2}$. Integrating this equation we have $G(t) = -\arctan t + c$ where c is the integration constant. However, from the integral expression for G we have $\lim_{t\to+\infty} G(t) = 0$, so $c = \frac{\pi}{2}$. Therefore $G(0) = \frac{\pi}{2}$. Q.E.D.

REMARK: We can use this construction to compute some other important integrals. Consider the integral $\int_0^\infty \frac{\sin x}{x} e^{-i\lambda x} dx$ for λ a real number. Taking real and imaginary parts of this expression and using the trigonometric identities

$$\sin\alpha \sin\beta = \frac{1}{2}[\cos(\alpha-\beta)-\cos(\alpha+\beta)], \quad \sin\alpha\cos\beta = \frac{1}{2}[\sin(\alpha-\beta)+\sin(\alpha+\beta)],$$

we can mimic the construction of the lemma to show that the improper Riemann integral converges for $|\lambda| \neq 1$. Then

$$\int_0^\infty \frac{\sin x}{x} e^{i\lambda x} dx = \lim_{\epsilon\to 0, \epsilon>0} G(i\lambda + \epsilon)$$

where, as we have shown, $G(t) = -\arctan t + \frac{\pi}{2} = \frac{i}{2} \ln \frac{1+it}{1-it} + \frac{\pi}{2}$. Using the property that $\ln(e^{i\theta}) = i\theta$ for $-\pi \leq \theta < \pi$ and taking the limit carefully, we find

$$\int_0^\infty \frac{\sin x}{x} e^{-i\lambda x} dx = \begin{cases} \frac{\pi}{2} + \frac{i}{2} \ln |\frac{1-\lambda}{1+\lambda}| & \text{for} |\lambda| < 1 \\ \frac{i}{2} \ln |\frac{1-\lambda}{1+\lambda}| & \text{for} |\lambda| > 1. \end{cases}$$

Noting that $\int_{-\infty}^\infty \frac{\sin x}{x} e^{-i\lambda x} dx = 2\int_0^\infty \frac{\sin x}{x} \cos(\lambda x) dx$ we find

$$\int_{-\infty}^\infty \frac{\sin x}{x} e^{-i\lambda x} dx = \begin{cases} \pi & \text{for} |\lambda| < 1 \\ \frac{\pi}{2} & \text{for} |\lambda| = 1 \\ 0 & \text{for} |\lambda| > 1 \end{cases} \tag{3.15}$$

Lemma 17 *Let $\delta > 0$, (think of δ as small) and $F(x)$ a function on $[0, \delta]$. Suppose*

- *$F(x)$ is piecewise continuous on $[0, \delta]$*
- *$F'(x)$ is piecewise continuous on $[0, \delta]$*
- *$F'(+0)$ exists.*

Then

$$\lim_{k\to\infty} \int_0^\delta \frac{\sin kx}{x} F(x) dx = \frac{\pi}{2} F(+0).$$

PROOF: We write

$$\int_0^\delta \frac{\sin kx}{x} F(x) dx = F(+0) \int_0^\delta \frac{\sin kx}{x} dx + \int_0^\delta \frac{F(x) - F(+0)}{x} \sin kx \, dx.$$

Set $G(x) = \frac{F(x)-F(+0)}{x}$ for $x \in [0, \delta]$ and $G(x) = 0$ elsewhere. Since $F'(+0)$ exists it follows that $G \in L^2$. hence, by the Riemann-Lebesgue Lemma, the second integral goes to 0 in the limit as $k \to \infty$. Hence

$$\lim_{k\to\infty}\int_0^\delta \frac{\sin kx}{x}F(x)dx = F(+0)\lim_{k\to\infty}\int_0^\delta \frac{\sin kx}{x}dx$$

$$= F(+0)\lim_{k\to\infty}\int_0^{k\delta}\frac{\sin u}{u}du = \frac{\pi}{2}F(+0).$$

For the last equality we have used our evaluation (3.13) of the integral of the sinc function. Q.E.D.

It is easy to compute the L^2 norm of sinc x:

Lemma 18

$$\int_0^\infty \frac{\sin^2 x}{x^2}dx = \pi\int_0^\infty \mathrm{sinc}^2 x\ dx = \frac{\pi}{2}. \tag{3.16}$$

PROOF: Integrate by parts.

$$\int_0^\infty \frac{\sin^2 x}{x^2}dx = -\frac{\sin^2 x}{x}|_0^\infty + \int_0^\infty \frac{2\sin x\cos x}{x}dx = \int_0^\infty \frac{\sin 2x}{x}dx$$

$$= \int_0^\infty \frac{\sin y}{y}dy = \frac{\pi}{2}.$$

Q.E.D.

Here is a more complicated proof, using the same technique as for Lemma 13. Set $G(t) = \int_0^\infty e^{-tx}(\frac{\sin x}{x})^2dx$, defined for $t \geq 0$. Now, $G''(t) = \int_0^\infty e^{-tx}\sin^2 x\ dx$ for $t > 0$. Integrating by parts we have

$$\int_0^\infty e^{-tx}\sin^2 x\ dx = 2\int_0^\infty \frac{e^{-tx}}{t^2}dx - 4\int_0^\infty \frac{e^{-tx}\sin x}{t^2}dx.$$

Hence, $G''(t) = \frac{2}{4t+t^3}$. Integrating this equation twice we have

$$G(t) = \frac{1}{2}t\ln t - \frac{1}{4}t\ln(4+t^2) - \arctan\frac{t}{2} + bt + c$$

where b, c are the integration constants. However, from the integral expression for G we have $\lim_{t\to+\infty} G(t) = \lim_{t\to+\infty} G'(t) = 0$, so $b = 0, c = \frac{\pi}{2}$. Therefore $G(0) = \frac{\pi}{2}$. Q.E.D.

Again we can use this construction to compute some other important integrals. Consider the integral $\int_0^\infty (\frac{\sin x}{x})^2 e^{-i\lambda x} dx$ for λ a real number. Then

$$\int_0^\infty (\frac{\sin x}{x})^2 e^{i\lambda x} dx = \lim_{\epsilon \to 0, \epsilon > 0} G(i\lambda + \epsilon)$$

where,

$$G(t) = \frac{1}{2} t \ln t - \frac{1}{4} t \ln(4 + t^2) - \arctan \frac{t}{2} + \frac{\pi}{2}.$$

Using the property that $\ln(e^{i\theta}) = i\theta$ for $-\pi \le \theta < \pi$ and taking the limit carefully, we find

$$\int_0^\infty (\frac{\sin x}{x})^2 e^{-i\lambda x} dx = \begin{cases} \frac{i\lambda}{2}(\ln|\lambda| \pm i\frac{\pi}{2}) - \frac{i\lambda}{4}\ln|4-\lambda^2| + \frac{i}{2}\ln|\frac{2-\lambda}{2+\lambda}| + \frac{\pi}{2} & \text{for} \pm \lambda \le 2 \\ \frac{i\lambda}{2}(\ln|\lambda| \pm i\frac{\pi}{2}) - \frac{i\lambda}{4}(\ln|4-\lambda^2| \pm i\pi) + \frac{i}{2}\ln|\frac{2-\lambda}{2+\lambda}| & \text{for} \pm \lambda > 2. \end{cases}$$

Noting that $\int_{-\infty}^\infty (\frac{\sin x}{x})^2 e^{-i\lambda x} dx = 2\int_0^\infty (\frac{\sin x}{x})^2 \cos(\lambda x) dx$ we find

$$\int_{-\infty}^\infty (\frac{\sin x}{x})^2 e^{-i\lambda x} dx = \begin{cases} \pi(1 - \frac{|\lambda|}{2}) & \text{for}|\lambda| < 2 \\ 0 & \text{for}|\lambda| \ge 2 \end{cases} \qquad (3.17)$$

3.4.3 The convergence proof: part 2

We return to the proof of the pointwise convergence theorem.

Theorem 21 *Suppose*

- *$f(t)$ is periodic with period 2π.*
- *$f(t)$ is piecewise continuous on $[0, 2\pi]$.*
- *$f'(t)$ is piecewise continuous on $[0, 2\pi]$.*

Then the Fourier series of $f(t)$ converges to $\frac{f(t+0)+f(t-0)}{2}$ at each point t.

END OF PROOF: We have

$$S_k(t) = \frac{a_0}{2} + \sum_{n=1}^{k} (a_n \cos nt + b_n \sin nt)$$

and

$$\lim_{k\to\infty} S_k(t) = \lim_{k\to\infty} \frac{1}{\pi} \int_0^\delta \frac{\sin[(k+\frac{1}{2})x]}{x} \frac{[f(t+x) + f(t-x)]x}{2\sin\frac{x}{2}} dx$$

$$= \lim_{k\to\infty} \frac{1}{\pi}\int_0^{\delta} \frac{\sin[(k+\frac{1}{2})x]}{x} F(x)dx$$

$$= \frac{F(+0)}{\pi} \lim_{k\to\infty} \int_0^{k\delta} \frac{\sin u}{u} du = \frac{\pi}{2\pi} F(+0)$$

by the last lemma. But $F(+0) = f(t+0) + f(t-0)$. Hence

$$\lim_{k\to\infty} S_k(t) = \frac{f(t+0) + f(t-0)}{2}.$$

Q.E.D.

3.4.4 An alternate (slick) pointwise convergence proof

We make the same assumptions about $f(t)$ as in the theorem above, and in addition we modify f, if necessary, so that

$$f(t) = \frac{f(t+0) + f(t-0)}{2}$$

at each point t. This condition affects the definition of f only at a finite number of points of discontinuity. It doesn't change any integrals and the values of the Fourier coefficients.

Lemma 19

$$\int_0^{2\pi} D_k(x)dx = \pi$$

PROOF:

$$\int_0^{2\pi} D_k(x)dx = \int_0^{2\pi} (\frac{1}{2} + \sum_{n=1}^{k} \cos nx)dx = \pi.$$

Q.E.D.

Using the Lemma we can write

$$S_k(t) - f(t) = \frac{1}{\pi}\int_0^{2\pi} D_k(t-x)[f(x) - f(t)]dx$$

$$= \frac{1}{\pi}\int_0^{\pi} D_k(x)[f(t+x) + f(t-x) - 2f(t)]dx$$

$$= \frac{1}{\pi}\int_0^{\pi} \frac{[f(t+x) + f(t-x) - 2f(t)]}{2\sin\frac{x}{2}} \sin(k+\frac{1}{2})x \; dx$$

$$= \frac{1}{\pi}\int_0^{\pi}[H_1(t,x)\sin kx + H_2(t,x)\cos kx]dx$$

¿From the assumptions, H_1, H_2 are square integrable in x. Indeed, we can use L'Hospital's rule and the assumptions that f and f' are piecewise continuous to show that the limit

$$\lim_{x\to 0}\frac{[f(t+x)+f(t-x)-2f(t)]}{2\sin\frac{x}{2}}$$

exists. Thus H_1, H_2 are bounded for $x = 0$. Then, by the Riemann-Lebesgue Lemma, the last expression goes to 0 as $k \to \infty$:

$$\lim_{k\to\infty}[S_k(t) - f(t)] = 0$$

.

3.4.5 Uniform pointwise convergence

We have shown that for functions f with the properties:

- $f(t)$ is periodic with period 2π.
- $f(t)$ is piecewise continuous on $[0, 2\pi]$.
- $f'(t)$ is piecewise continuous on $[0, 2\pi]$.

then at each point t the partial sums of the Fourier series of f,

$$\begin{aligned} f(t) &\sim \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n\cos nt + b_n\sin nt) = S(t), \qquad a_n = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos nt\, dt \\ b_n &= \frac{1}{\pi}\int_0^{2\pi} f(t)\sin nt\, dt, \end{aligned}$$

converge to $\frac{f(t+0)+f(t-0)}{2}$:

$$S_k(t) = \frac{a_0}{2} + \sum_{n=1}^{k}(a_n\cos nt + b_n\sin nt),$$

$$\lim_{k\to\infty} S_k(t) = \frac{f(t+0)+f(t-0)}{2}.$$

(If we require that f satisfies $f(t) = \frac{f(t+0)+f(t-0)}{2}$ at each point then the series will converge to f everywhere. In this section I will make this requirement.) Now we want to examine the rate of convergence.

We know that for every $\epsilon > 0$ we can find an integer $N(\epsilon, t)$ such that $|S_k(t) - f(t)| < \epsilon$ for every $k > N(\epsilon, t)|$. Then the finite sum trigonometric polynomial $S_k(t)$ will approximate $f(t)$ with an error $< \epsilon$. However, in general N depends on the point t; we have to recompute it for each t. What we would prefer is *uniform convergence*. The Fourier series of f will converge to f *uniformly* if for every $\epsilon > 0$ we can find an integer $N(\epsilon)$ such that $|S_k(t) - f(t)| < \epsilon$ for every $k > N(\epsilon)$ and *for all* t. Then the finite sum trigonometric polynomial $S_k(t)$ will approximate $f(t)$ everywhere with an error $< \epsilon$.

We cannot achieve uniform convergence for all functions f in the class above. The partial sums are continuous functions of t, Recall from calculus that if a sequence of continuous functions converges uniformly, the limit function is also continuous. Thus for any function f with discontinuities, we cannot have uniform convergence of the Fourier series.

If f is continuous, however, then we do have uniform convergence.

Theorem 22 *Assume f has the properties:*

- *$f(t)$ is periodic with period 2π.*
- *$f(t)$ is continuous on $[0, 2\pi]$.*
- *$f'(t)$ is piecewise continuous on $[0, 2\pi]$.*

Then the Fourier series of f converges uniformly.

PROOF: Consider the Fourier series of both f and f':

$$\begin{aligned} f(t) &\sim \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt), \qquad a_n = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos nt\, dt \\ b_n &= \frac{1}{\pi}\int_0^{2\pi} f(t)\sin nt\, dt, \end{aligned}$$

$$\begin{aligned} f'(t) &\sim \frac{A_0}{2} + \sum_{n=1}^{\infty}(A_n \cos nt + B_n \sin nt), \qquad A_n = \frac{1}{\pi}\int_0^{2\pi} f'(t)\cos nt\, dt \\ B_n &= \frac{1}{\pi}\int_0^{2\pi} f'(t)\sin nt\, dt, \end{aligned}$$

Now

$$A_n = \frac{1}{\pi}\int_0^{2\pi} f'(t)\cos nt\; dt = \frac{1}{\pi} f(t)\cos nt|_0^{2\pi} + \frac{n}{\pi}\int_0^{2\pi} f(t)\sin nt\; dt$$

$$= \begin{cases} nb_n, & n \geq 1 \\ 0, & n = 0 \end{cases}$$

(We have used the fact that $f(0) = f(2\pi)$.) Similarly,

$$B_n = \frac{1}{\pi}\int_0^{2\pi} f'(t)\sin nt\; dt = \frac{1}{\pi} f(t)\sin nt|_0^{2\pi} - \frac{n}{\pi}\int_0^{2\pi} f(t)\cos nt\; dt = -na_n.$$

Using Bessel's inequality for f' we have

$$\sum_{n=1}^{\infty}(|A_n|^2 + |B_n|^2) \leq \frac{1}{\pi}\int_0^{2\pi} |f'(t)|^2 dt < \infty,$$

hence

$$\sum_{n=1}^{\infty} n^2(|a_n|^2 + |b_n|^2) \leq \infty.$$

Now

$$\begin{matrix} \sum_{n=1}^m |a_n| \\ \sum_{n=1}^m |b_n| \end{matrix} \leq \sum_{n=1}^m \sqrt{|a_n|^2 + |b_n|^2} = \sum_{n=1}^m \frac{1}{n}\sqrt{|A_n|^2 + |B_n|^2}$$

$$\leq (\sum_{n=1}^m \frac{1}{n^2})(\sum_{n=1}^m (|A_n|^2 + |B_n|^2))$$

which converges as $m \to \infty$. (We have used the Schwarz inequality for the last step.) Hence $\sum_{n=1}^\infty |a_n| < \infty$, $\sum_{n=1}^\infty |b_n| < \infty$. Now

$$|\frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt)| \leq |\frac{a_0}{2}| + \sum_{n=1}^{\infty}(|a_n \cos nt| + |b_n \sin nt|) \leq$$

$$|\frac{a_0}{2}| + \sum_{n=1}^{\infty}(|a_n| + |b_n|) < \infty,$$

so the series converges uniformly and absolutely. Q.E.D.

Corollary 7 *Parseval's Theorem. For f satisfying the hypotheses of the preceding theorem*

$$\frac{|a_0|^2}{2} + \sum_{n=1}^{\infty}(|a_n|^2 + |b_n|^2) = \frac{1}{\pi}\int_0^{2\pi} |f(t)|^2 dt.$$

PROOF: The Fourier series of f converges uniformly: for any $\epsilon > 0$ there is an integer $N(\epsilon)$ such that $|S_k(t) - f(t)| < \epsilon$ for every $k > N(\epsilon)$ and *for all* t. Thus

$$\int_0^{2\pi} |S_k(t) - f(t)|^2 \, dt = ||S_k - f||^2 = ||f||^2 - \pi(\frac{|a_0|^2}{2} + \sum_{n=1}^{k}(|a_n|^2 + |b_n|^2) < \frac{\epsilon^2}{2\pi}$$

for $k > N(\epsilon)$. Q.E.D.

REMARK 1: Parseval's Theorem actually holds for any $f \in L^2[0, 2\pi]$, as we shall show later.

REMARK 2: As the proof of the preceding theorem illustrates, differentiability of a function improves convergence of its Fourier series. The more derivatives the faster the convergence. There are famous examples to show that continuity alone is *not* sufficient for convergence.

3.5 More on pointwise convergence, Gibbs phenomena

Let's return to our Example 1 of Fourier series:

$$h(t) = \begin{cases} 0, & t = 0 \\ \frac{\pi - t}{2}, & 0 < t < 2\pi \\ 0, & t = 2\pi. \end{cases}$$

and $h(t + 2\pi) = h(t)$. In this case, $a_n = 0$ for all n and $b_n = \frac{1}{n}$. Therefore,

$$\frac{\pi - t}{2} = \sum_{n=1}^{\infty} \frac{\sin nt}{n}, \qquad 0 < t < 2\pi.$$

The function h has a discontinuity at $t = 0$ so the convergence of this series can't be uniform. Let's examine this case carefully. What happens to the partial sums near the discontinuity?

Here, $S_k(t) = \sum_{n=1}^{k} \frac{\sin nt}{n}$ so

$$S'_k(t) = \sum_{n=1}^{k} \cos nt = D_k(t) - \frac{1}{2} = \frac{\sin(k + \frac{1}{2})t}{2 \sin \frac{t}{2}} - \frac{1}{2} = \frac{\sin \frac{kt}{2} \cos \frac{(k+1)t}{2}}{\sin \frac{t}{2}}.$$

Thus, since $S_k(0) = 0$ we have

$$S_k(t) = \int_0^t S'_k(x) dx = \int_0^t \left(\frac{\sin \frac{kx}{2} \cos \frac{(k+1)x}{2}}{2 \sin \frac{x}{2}} \right) dx.$$

Note that $S'_k(0) > 0$ so that S_k starts out at 0 for $t = 0$ and then increases. Looking at the derivative of S_k we see that the first maximum is at the critical point $t_k = \frac{\pi}{k+1}$ (the first zero of $\cos\frac{(k+1)x}{2}$ as x increases from 0). Here, $h(t_k) = \frac{\pi - t_k}{2}$. The error is

$$S_k(t_k) - h(t_k) = \int_0^{t_k} \frac{\sin(k+\frac{1}{2})x}{2\sin\frac{x}{2}}dx - \frac{\pi}{2}$$

$$= \int_0^{t_k} \frac{\sin(k+\frac{1}{2})x}{x}dx + \int_0^{t_k}\left(\frac{1}{2\sin\frac{x}{2}} - \frac{1}{x}\right)\sin(k+\frac{1}{2})x\, dx - \frac{\pi}{2}$$

$$= I(t_k) + J(t_k) - \frac{\pi}{2}$$

where

$$I(t_k) = \int_0^{t_k} \frac{\sin(k+\frac{1}{2})x}{x}dx = \int_0^{(k+\frac{1}{2})t_k} \frac{\sin u}{u}du \to \int_0^{\pi} \frac{\sin u}{u}du \approx 1.851397052$$

(according to MAPLE). Also

$$J(t_k) = \int_0^{t_k}\left(\frac{1}{2\sin\frac{x}{2}} - \frac{1}{x}\right)[\sin kx \cos\frac{x}{2} + \cos nx \sin\frac{x}{2}]\, dx.$$

Note that the quantity in the round braces is bounded near $x = 0$, hence by the Riemann-Lebesgue Lemma we have $J(t_k) \to 0$ as $k \to \infty$. We conclude that

$$\lim_{k\to\infty}(S_k(t_k) - h(t_k)) \approx 1.851397052 - \frac{\pi}{2} \approx .280600725$$

To sum up,$\lim_{k\to\infty} S_k(t_k) \approx 1.851397052$ whereas $\lim_{k\to\infty} h(t_k) = \frac{\pi}{2} \approx 1.570796327$. The partial sum is overshooting the correct value by about 17.86359697%! This is called the *Gibbs Phenomenon*.

To understand it we need to look more carefully at the convergence properties of the partial sums $S_k(t) = \sum_{n=1}^k \frac{\sin nt}{n}$ for all t.

First some preparatory calculations. Consider the geometric series $E_k(t) = \sum_{n=1}^k e^{int} = \frac{e^{it}(1-e^{ikt})}{1-e^{it}}$.

Lemma 20 *For* $0 < t < 2\pi$,

$$|E_k(t)| \le \frac{2}{|1-e^{it}|} = \frac{1}{\sin\frac{t}{2}}.$$

Note that $S_k(t)$ is the imaginary part of the complex series $\sum_{n=1}^k \frac{e^{int}}{n}$.

Lemma 21 *Let* $0 < \alpha < \beta < 2\pi$*. The series* $\sum_{n=1}^{\infty} \frac{e^{int}}{n}$ *converges uniformly for all* t *in the interval* $[\alpha, \beta]$*.*

PROOF: (tricky)

$$\sum_{n=j}^{k} \frac{e^{int}}{n} = \sum_{n=j}^{k} \frac{E_n(t) - E_{n-1}(t)}{n} = \sum_{n=j}^{k} \frac{E_n(t)}{n} - \sum_{n=j}^{k} \frac{E_n(t)}{n+1} - \frac{E_{j-1}(t)}{j} + \frac{E_k(t)}{K+1}$$

and

$$|\sum_{n=j}^{k} \frac{e^{int}}{n}| \leq \frac{1}{\sin\frac{t}{2}} \left(\sum_{n=j}^{k} (\frac{1}{n} - \frac{1}{n+1}) + \frac{1}{j} + \frac{1}{k+1} \right) = \frac{2}{j \sin\frac{t}{2}}.$$

This implies by the Cauchy Criterion that $\sum_{n=j}^{k} \frac{e^{int}}{n}$ converges uniformly on $[\alpha, \beta]$. Q.E.D.

This shows that the Fourier series for $h(t)$ converges uniformly on any closed interval that doesn't contain the discontinuities at $t = 2\pi\ell$, $\ell = 0, \pm 1, \pm 2, \cdots$. Next we will show that the partial sums $S_k(t)$ are bounded for *all* t and all k. Thus, even though there is an overshoot near the discontinuities, the overshoot is strictly bounded.

¿From the lemma on uniform convergence above we already know that the partial sums are bounded on any closed interval not containing a discontinuity. Also, $S_k(0) = 0$ and $S_k(-t) = -S_k(t)$, so it suffices to consider the interval $0 < t < \frac{\pi}{2}$.

We will use the facts that $\frac{2}{\pi} \leq \frac{\sin t}{t} \leq 1$ for $0 < t \leq \frac{\pi}{2}$. The right-hand inequality is a basic calculus fact and the left-hand one is obtained by solving the calculus problem of minimizing $\frac{\sin t}{t}$ over the interval $0 < t < \frac{\pi}{2}$. Note that

$$|\sum_{n=1}^{k} \frac{\sin nt}{n}| \leq |\sum_{1 \leq n < 1/t} \frac{t \sin nt}{nt}| + |\sum_{1/t \leq n \leq k} \frac{\sin nt}{n}|.$$

Using the calculus inequalities and the lemma, we have

$$|\sum_{n=1}^{k} \frac{\sin nt}{n}| \leq \sum_{1 \leq n < 1/t} t + \frac{2}{\frac{1}{t} \sin\frac{t}{2}} \leq t \cdot \frac{1}{t} + \frac{2}{\frac{1}{t}\frac{t}{2}\frac{2}{\pi}} = 1 + 2\pi.$$

Thus the partial sums are uniformly bounded for all t and all k.

We conclude that the Fourier series for $h(t)$ converges uniformly to $h(t)$ in any closed interval not including a discontinuity. Furthermore the partial sums of the Fourier series are uniformly bounded. At each discontinuity $t_N = 2\pi N$ of h the partial sums S_k overshoot $h(t_N + 0)$ by about 17.9% (approaching from the right) as $k \to \infty$ and undershoot $h(t_N - 0)$ by the same amount.

All of the work that we have put into this single example will pay off, because the facts that have emerged are of broad validity. Indeed we can consider any function f satisfying our usual conditions as the sum of a continuous function for which the convergence is uniform everywhere and a finite number of translated and scaled copies of $h(t)$.

Theorem 23 *Let f be a complex valued function such that*

- *$f(t)$ is periodic with period 2π.*
- *$f(t)$ is piecewise continuous on $[0, 2\pi]$.*
- *$f'(t)$ is piecewise continuous on $[0, 2\pi]$.*
- *$f(t) = \frac{f(t+0)+f(t-0)}{2}$ at each point t.*

Then

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt)$$

pointwise. The convergence of the series is uniform on every closed interval in which f is continuous.

PROOF: let $x_1, x_2, \cdots, x_\ell$ be the points of discontinuity of f in $[0, 2\pi)$ Set $s(x_j) = f(x_j + 0) - f(x_j - 0)$. Then the function

$$g(t) = f(t) - \sum_{j=1}^{\ell} \frac{s(x_j)}{\pi} h(t - x_j)$$

is everywhere continuous and also satisfies all of the hypotheses of the theorem. Indeed, at the discontinuity x_j of f we have

$$g(x_j - 0) = f(x_j - 0) - \frac{2s(x_j)}{\pi} h(-0) = f(x_j - 0) - \frac{f(x_j + 0) - f(x_j - 0))}{\pi}(\frac{-\pi}{2})$$

$$= \frac{f(x_j - 0) + f(x_j + 0)}{2} = f(x_j).$$

Similarly, $g(x_j + 0) = f(x_j)$. Therefore $g(t)$ can be expanded in a Fourier series that converges absolutely and uniformly. However, $\sum_{j=1}^{\ell} \frac{s(x_j)}{\pi} h(t - x_j)$ can be expanded in a Fourier series that converges pointwise and uniformly in every closed interval that doesn't include a discontinuity. But

$$f(t) = g(t) + \sum_{j=1}^{\ell} \frac{s(x_j)}{\pi} h(t - x_j),$$

and the conclusion follows. Q.E.D.

Corollary 8 *Parseval's Theorem. For f satisfying the hypotheses of the preceding theorem*

$$\frac{|a_0|^2}{2} + \sum_{n=1}^{\infty} (|a_n|^2 + |b_n|^2) = \frac{1}{\pi} \int_0^{2\pi} |f(t)|^2 dt.$$

PROOF: As in the proof of the theorem, let $x_1, x_2, \cdots, x_\ell$ be the points of discontinuity of f in $[0, 2\pi)$ and set $s(x_j) = f(x_j + 0) - f(x_j - 0)$. Choose $a \geq 0$ such that the discontinuities are contained in the interior of $I = (-a, 2\pi - a)$. ¿From our earlier results we know that the partial sums of the Fourier series of h are uniformly bounded with bound $M > 0$. Choose $P = \sup_{t \in [0, 2\pi]} |f(t)|$. Then $|S_k(t) - f(t)|^2 \leq (M + P)^2$ for all t and all k. Given $\epsilon > 0$ choose non-overlapping open intervals $I_1, I_2, \cdots, I_\ell \subset I$ such that $x_j \in I_j$ and $(\sum_{j=1}^{\ell} |I_j|)(M + P)^2 < \frac{\epsilon}{2}$. Here, $|I_j|$ is the length of the interval I_j. Now the Fourier series of f converges uniformly on the closed set $A = [-a, 2\pi - a] - I_1 \cup I_2 \cup \cdots \cup I_\ell$. Choose an integer $N(\epsilon)$ such that $|S_k(t) - f(t)|^2 < \frac{\epsilon}{4\pi}$ for all $t \in A, k \geq N(\epsilon)$. Then

$$\int_0^{2\pi} |S_k(t) - f(t)|^2 dt = \int_{-a}^{2\pi - a} |S_k(t) - f(t)|^2 dt =$$

$$\int_A |S_k(t) - f(t)|^2 dt + \int_{I_1 \cup I_2 \cup \cdots \cup I_\ell} |S_k(t) - f(t)|^2 dt$$

$$< 2\pi \frac{\epsilon}{4\pi} + (\sum_{j=1}^{\ell} |I_j|)(M + P)^2 < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Thus $\lim_{k \to \infty} ||S_k - f|| = 0$ and the partial sums converge to f in the mean.

Furthermore,

$$\epsilon > \int_0^{2\pi} |S_k(t) - f(t)|^2 \, dt = ||S_k - f||^2 = ||f||^2 - \pi(\frac{|a_0|^2}{2} + \sum_{n=1}^{k} (|a_n|^2 + |b_n|^2)$$

for $k > N(\epsilon)$. Q.E.D.

3.6 Mean convergence, Parseval's equality, Integration and differentiation of Fourier series

The convergence theorem and the version of the Parseval identity proved immediately above apply to step functions on $[0, 2\pi]$. However, we already know that the space of step functions on $[0, 2\pi]$ is dense in $L^2[0, 2\pi]$. Since every step function is the limit in the norm of the partial sums of its Fourier series, this means that the space of all finite linear combinations of the functions $\{e^{int}\}$ is dense in $L^2[0, 2\pi]$. Hence $\{e^{int}/\sqrt{2\pi}\}$ is an ON basis for $L^2[0, 2\pi]$ and we have the

Theorem 24 *Parseval's Equality (strong form) [Plancherel Theorem]. If $f \in L^2[0, 2\pi]$*

$$\frac{|a_0|^2}{2} + \sum_{n=1}^{\infty}(|a_n|^2 + |b_n|^2) = \frac{1}{\pi}\int_0^{2\pi} |f(t)|^2 dt,$$

where a_n, b_n are the Fourier coefficients of f.

Integration of a Fourier series term-by-term yields a series with improved convergence.

Theorem 25 *Let f be a complex valued function such that*

- *$f(t)$ is periodic with period 2π.*
- *$f(t)$ is piecewise continuous on $[0, 2\pi]$.*
- *$f'(t)$ is piecewise continuous on $[0, 2\pi]$.*

Let

$$f(t) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt)$$

be the Fourier series of f. Then

$$\int_0^t f(x)dx = \frac{a_0}{2}t + \sum_{n=1}^{\infty} \frac{1}{n}\left[a_n \sin nt - b_n(\cos nt - 1)\right].$$

where the convergence is uniform on the interval $[0, 2\pi]$.

PROOF: Let $F(t) = \int_0^t f(x)dx - \frac{a_0}{2}t$. Then

- $F(2\pi) = \int_0^{2\pi} f(x)dx - \frac{a_0}{2}(2\pi) = 0 = F(0)$.

- $F(t)$ is continuous on $[0, 2\pi]$.
- $F'(t) = f(t) - \frac{a_0}{2}$ is piecewise continuous on $[0, 2\pi]$.

Thus the Fourier series of F converges to F uniformly and absolutely on $[0, 2\pi]$:

$$F(t) = \frac{A_0}{2} + \sum_{n=1}^{\infty}(A_n \cos nt + B_n \sin nt).$$

Now

$$A_n = \frac{1}{\pi}\int_0^{2\pi} F(t)\cos nt\, dt = \frac{F(t)\sin nt}{n\pi}|_0^{2\pi} - \frac{1}{n\pi}\int_0^{2\pi}(f(t) - \frac{a_0}{2})\sin nt\, dt$$
$$= -\frac{b_n}{n}, \qquad n \neq 0,$$

and

$$B_n = \frac{1}{\pi}\int_0^{2\pi} F(t)\sin nt\, dt = -\frac{F(t)\cos nt}{n\pi}|_0^{2\pi} + \frac{1}{n\pi}\int_0^{2\pi}(f(t) - \frac{a_0}{2})\cos nt\, dt$$
$$= \frac{a_n}{n}.$$

Therefore,

$$F(t) = \frac{A_0}{2} + \sum_{n=1}^{\infty}(-\frac{b_n}{n}\cos nt + \frac{a_n}{n}\sin nt),$$
$$F(2\pi) = 0 = \frac{A_0}{2} - \sum_{n=1}^{\infty}\frac{b_n}{n}.$$

Solving for A_0 we find

$$F(t) = \int_0^t f(x)dx - \frac{a_0}{2}t = \sum_{n=1}^{\infty}\frac{1}{n}\left[a_n \sin nt - b_n(\cos nt - 1)\right].$$

Q.E.D.

Example 2 *Let*

$$f(t) = \begin{cases} \frac{\pi - t}{2} & 0 < t < 2\pi \\ 0 & t = 0, 2\pi. \end{cases}$$

Then

$$\frac{\pi - t}{2} \sim \sum_{n=1}^{\infty}\frac{\sin nt}{n}.$$

Integrating term-by term we find

$$\frac{2\pi t - t^2}{4} = -\sum_{n=1}^{\infty} \frac{1}{n^2}(\cos nt - 1), \quad 0 \le t \le 2\pi.$$

Differentiation of Fourier series, however, makes them less smooth and may not be allowed. For example, differentiating the Fourier series

$$\frac{\pi - t}{2} \sim \sum_{n=1}^{\infty} \frac{\sin nt}{n},$$

formally term-by term we get

$$-\frac{1}{2} \sim \sum_{n=1}^{\infty} \cos nt,$$

which doesn't converge on $[0, 2\pi]$. In fact it can't possible be a Fourier series for an element of $L^2[0, 2\pi]$. (Why?)

If f is sufficiently smooth and periodic it *is* OK to differentiate term-by-term to get a new Fourier series.

Theorem 26 *Let f be a complex valued function such that*

- *$f(t)$ is periodic with period 2π.*
- *$f(t)$ is continuous on $[0, 2\pi]$.*
- *$f'(t)$ is piecewise continuous on $[0, 2\pi]$.*
- *$f''(t)$ is piecewise continuous on $[0, 2\pi]$.*

Let

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt)$$

be the Fourier series of f. Then at each point $t \in [0, 2\pi]$ where $f''(t)$ exists we have

$$f'(t) = \sum_{n=1}^{\infty} n\left[-a_n \sin nt + b_n \cos nt\right].$$

PROOF: By the Fourier convergence theorem the Fourier series of f' converges to $\frac{f'(t_0+0)+f'(t_0-0)}{2}$ at each point t_0. If $f''(t_0)$ exists at the point then the Fourier series converges to $f'(t_0)$, where

$$f'(t) \sim \frac{A_0}{2} + \sum_{n=1}^{\infty}(A_n \cos nt + B_n \sin nt).$$

Now

$$A_n = \frac{1}{\pi}\int_0^{2\pi} f'(t)\cos nt\, dt = \frac{f(t)\cos nt}{\pi}|_0^{2\pi} + \frac{n}{\pi}\int_0^{2\pi} f(t)\sin nt\, dt$$

$$= nb_n,$$

$A_0 = \frac{1}{\pi}\int_0^{2\pi} f'(t)dt = \frac{1}{\pi}(f(2\pi) - f(0)) = 0$ (where, if necessary, we adjust the interval of length 2π so that f' is continuous at the endpoints) and

$$B_n = \frac{1}{\pi}\int_0^{2\pi} f'(t)\sin nt\, dt = \frac{f(t)\sin nt}{\pi}|_0^{2\pi} - \frac{n}{\pi}\int_0^{2\pi} f(t)\cos nt\, dt$$

$$= -na_n.$$

Therefore,

$$f'(t) \sim \sum_{n=1}^{\infty} n(b_n \cos nt - a_n \sin nt).$$

Q.E.D.

Note the importance of the requirement in the theorem that f is continuous everywhere and periodic, so that the boundary terms vanish in the integration by parts formulas for A_n and B_n. Thus it is OK to differentiate the Fourier series

$$f(t) = \frac{2\pi t - t^2}{4} - \frac{\pi^2}{6} = -\sum_{n=1}^{\infty}\frac{1}{n^2}\cos nt, \quad 0 \le t \le 2\pi$$

term-by term, where $f(0) = f(2\pi)$, to get

$$f'(t) = \frac{\pi - t}{2} \sim \sum_{n=1}^{\infty}\frac{\sin nt}{n}.$$

However, even though $f'(t)$ is infinitely differentiable for $0 < t < 2\pi$ we have $f'(0) \neq f'(2\pi)$, so we cannot differentiate the series again.

3.7 Arithmetic summability and Fejér's theorem

We know that the kth partial sum of the Fourier series of a square integrable function f:

$$S_k(t) = \frac{a_0}{2} + \sum_{n=1}^{k}(a_n \cos nt + b_n \sin nt)$$

is the trigonometric polynomial of order k that best approximates f in the Hilbert space sense. However, the limit of the partial sums

$$S(t) = \lim_{k\to\infty} S_k(t),$$

doesn't necessarily converge pointwise. We have proved pointwise convergence for piecewise smooth functions, but if, say, all we know is that f is continuous or then pointwise convergence is much harder to establish. Indeed there are examples of continuous functions whose Fourier series diverges at uncountably many points. Furthermore we have seen that at points of discontinuity the Gibbs phenomenon occurs and the partial sums overshoot the function values. In this section we will look at another way to recapture $f(t)$ from its Fourier coefficients, by Cesàro sums (arithmetic means). This method is surprisingly simple, gives uniform convergence for continuous functions $f(t)$ and avoids most of the Gibbs phenomena difficulties.

The basic idea is to use the arithmetic means of the partial sums to approximate f. Recall that the kth partial sum of $f(t)$ is defined by

$$S_k(t) = \frac{1}{2\pi}\int_0^{2\pi} f(x)dx + \frac{1}{\pi}\sum_{n=1}^{k}\left(\int_0^{2\pi} f(x)\cos nx\, dx \cos nt + \int_0^{2\pi} f(x)\sin nx\, dx \sin nt\right),$$

so

$$\begin{aligned} S_k(t) &= \frac{1}{\pi}\int_0^{2\pi}\left[\frac{1}{2} + \sum_{n=1}^{k}(\cos nx \cos nt + \sin nx \sin nt\right] f(x)dx \\ &= \frac{1}{\pi}\int_0^{2\pi}\left[\frac{1}{2} + \sum_{n=1}^{k}\cos[n(t-x)]\right] f(x)dx \\ &= \frac{1}{\pi}\int_0^{2\pi} D_k(t-x)f(x)dx. \end{aligned}$$

where the kernel $D_k(t) = \frac{1}{2} + \sum_{n=1}^{k}\cos nt = -\frac{1}{2} + \sum_{m=0}^{k}\cos mt$. Further,

$$D_k(t) = \frac{\cos kt - \cos(k+1)t}{4\sin^2\frac{t}{2}} = \frac{\sin(k+\frac{1}{2})t}{2\sin\frac{t}{2}}.$$

Rather than use the partial sums $S_k(t)$ to approximate $f(t)$ we use the arithmetic means $\sigma_k(t)$ of these partial sums:

$$\sigma_k(t) = \frac{S_0(t) + S_1(t) + \cdots + S_{k-1}(t)}{k}, \qquad k = 1, 2, \cdots. \tag{3.18}$$

Then we have

$$\sigma_k(t) = \frac{1}{k\pi}\sum_{j=0}^{k-1}\int_0^{2\pi} D_j(t-x)f(x)dx = \int_0^{2\pi}\left[\frac{1}{k\pi}\sum_{j=0}^{k-1} D_j(t-x)\right] f(x)dx$$

$$= \frac{1}{\pi}\int_0^{2\pi} F_k(t-x)f(x)dx \tag{3.19}$$

where

$$F_k(t) = \frac{1}{k}\sum_{j=0}^{k-1} D_j(t) = \frac{1}{k}\sum_{j=0}^{k-1}\frac{\sin(j+\frac{1}{2})t}{2\sin\frac{t}{2}}.$$

Lemma 22

$$F_k(t) = \frac{1}{k}\left(\frac{\sin kt/2}{\sin t/2}\right)^2.$$

PROOF: Using the geometric series, we have

$$\sum_{j=0}^{k-1} e^{i(j+\frac{1}{2})t} = e^{i\frac{t}{2}}\frac{e^{ikt}-1}{e^{it}-1} = e^{i\frac{kt}{2}}\frac{\sin\frac{kt}{2}}{\sin\frac{t}{2}}.$$

Taking the imaginary part of this identity we find

$$F_k(t) = \frac{1}{k\sin\frac{t}{2}}\sum_{j=0}^{k-1}\sin(j+\frac{1}{2})t = \frac{1}{k}\left(\frac{\sin kt/2}{\sin t/2}\right)^2.$$

Q.E.D.

Note that F has the properties:

- $F_k(t) = F_k(t+2\pi)$
- $F_k(-t) = F_k(t)$
- $F_k(t)$ is defined and differentiable for all t and $F_k(0) = k$

- $F_k(t) \geq 0$.

¿From these properties it follows that the integrand of (3.19) is a 2π-periodic function of x, so that we can take the integral over any full 2π-period. Finally, we can change variables and divide up the integral, in analogy with our study of the Fourier kernel $D_k(t)$, and obtain the following simple expression for the arithmetic means:

Lemma 23

$$\sigma_k(t) = \frac{2}{k\pi} \int_0^{\pi/2} \frac{f(t+2x) + f(t-2x)}{2} \left(\frac{\sin kx}{\sin x}\right)^2 dx.$$

Lemma 24

$$\frac{2}{k\pi} \int_0^{\pi/2} \left(\frac{\sin kx}{\sin x}\right)^2 dx = 1.$$

PROOF: Let $f(t) \equiv 1$ for all t. Then $\sigma_k(t) \equiv 1$ for all k and t. Substituting into the expression from lemma 23 we obtain the result. Q.E.D.

Theorem 27 *(Fejér) Suppose $f(t) \in L^1[0, 2\pi]$, periodic with period 2π and let*

$$\sigma(t) = \lim_{x \to 0+} \frac{f(t+x) + f(t-x)}{2} = \frac{f(t+0) + f(t-0)}{2}$$

whenever the limit exists. For any t such that $\sigma(t)$ is defined we have

$$\lim_{k \to \infty} \sigma_k(t) = \sigma(t) = \frac{f(t+0) + f(t-0)}{2}.$$

PROOF: From lemmas 23 and 24 we have

$$\sigma_k(t) - \sigma(t) = \frac{2}{k\pi} \int_0^{\pi/2} \left[\frac{f(t+2x) + f(t-2x)}{2} - \sigma(t)\right] \left(\frac{\sin kx}{\sin x}\right)^2 dx.$$

For any t for which $\sigma(t)$ is defined, let $G_t(x) = \frac{f(t+2x)+f(t-2x)}{2} - \sigma(t)$. Then $G_t(x) \to 0$ as $t \to 0$ through positive values. Thus, given $\epsilon > 0$ there is a $\delta < \pi/2$ such that $|G_t(x)| < \epsilon/2$ whenever $0 < x \leq \delta$. We have

$$\sigma_k(t) - \sigma(t) = \frac{2}{k\pi} \int_0^{\delta} G_t(x) \left(\frac{\sin kx}{\sin x}\right)^2 dx + \frac{2}{k\pi} \int_{\delta}^{\pi/2} G_t(x) \left(\frac{\sin kx}{\sin x}\right)^2 dx.$$

Now

$$\left|\frac{2}{k\pi}\int_0^{\delta} G_t(x)\left(\frac{\sin kx}{\sin x}\right)^2 dx\right| \leq \frac{\epsilon}{k\pi}\int_0^{\pi/2}\left(\frac{\sin kx}{\sin x}\right)^2 dx = \frac{\epsilon}{2}$$

and

$$\left|\frac{2}{k\pi}\int_{\delta}^{\pi/2} G_t(x)\left(\frac{\sin kx}{\sin x}\right)^2 dx\right| \leq \frac{2}{k\pi\sin^2\delta}\int_{\delta}^{\pi/2}|G_t(x)|dx \leq \frac{2I}{k\pi\sin^2\delta},$$

where $I = \int_0^{\pi/2}|G_t(x)|dx$. This last integral exists because F is in L^1. Now choose K so large that $2I/(N\pi\sin^2\delta) < \epsilon/2$. Then if $k \geq K$ we have

$$|\sigma_k(t) - \sigma(t)\left| \leq |\frac{2}{k\pi}\int_0^{\delta} G_t(x)\left(\frac{\sin kx}{\sin x}\right)^2 dx\right| + \left|\frac{2}{k\pi}\int_{\delta}^{\pi/2} G_t(x)\left(\frac{\sin kx}{\sin x}\right)^2 dx\right| < \epsilon.$$

Q.E.D.

Corollary 9 *Suppose $f(t)$ satisfies the hypotheses of the theorem and also is continuous on the closed interval $[a,b]$. Then the sequence of arithmetic means $\sigma_k(t)$ converges uniformly to $f(t)$ on $[a,b]$.*

PROOF: If f is continuous on the closed bounded interval $[a,b]$ then it is uniformly continuous on that interval and the function G_t is bounded on $[a,b]$ with upper bound M, independent of t. Furthermore one can determine the δ in the preceding theorem so that $|G_t(x)| < \epsilon/2$ whenever $0 < x \leq \delta$ and uniformly for all $t \in [a,b]$. Thus we can conclude that $\sigma_k \to \sigma$, uniformly on $[a,b]$. Since f is continuous on $[a,b]$ we have $\sigma(t) = f(t)$ for all $t \in [a,b]$. Q.E.D.

Corollary 10 *(Weierstrass approximation theorem) Suppose $f(t)$ is real and continuous on the closed interval $[a,b]$. Then for any $\epsilon > 0$ there exists a polynomial $p(t)$ such that*

$$|f(t) - p(t)| < \epsilon$$

for every $t \in [a,b]$.

SKETCH OF PROOF: Using the methods of Section 3.3 we can find a linear transformation to map $[a,b]$ one-to-one on a closed subinterval $[a',b']$ of $[0,2\pi]$, such that $0 < a' < b' < 2\pi$. This transformation will take polynomials in t to polynomials. Thus, without loss of generality, we can assume $0 < a < b < 2\pi$.

Let $g(t) = f(t)$ for $a \le t \le b$ and define $g(t)$ outside that interval so that it is continuous at $T = a, b$ and is periodic with period 2π. Then from the first corollary to Fejér's theorem, given an $\epsilon > 0$ there is an integer N and arithmetic sum

$$\sigma(t) = \frac{A_0}{2} + \sum_{j=1}^{N} (A_j \cos jt + B_j \sin jt)$$

such that $|f(t) - \sigma(t)| = |g(t) - \sigma(t)| < \frac{\epsilon}{2}$ for $a \le t \le b$. Now $\sigma(t)$ is a trigonometric polynomial and it determines a poser series expansion in t about the origin that converges uniformly on every finite interval. The partial sums of this power series determine a series of polynomials $\{p_n(t)\}$ of order n such that $p_n \to \sigma$ uniformly on $[a, b]$, Thus there is an M such that $|\sigma(t) - p_M(t)| < \frac{\epsilon}{2}$ for all $t \in [a, b]$. Thus

$$|f(t) - p_M(t)| \le |f(t) - \sigma(t)| + |\sigma(t) - p_M(t)| < \epsilon$$

for all $t \in [a, b]$. Q.E.D.

This important result implies not only that a continuous function on a bounded interval can be approximated uniformly by a polynomial function but also (since the convergence is uniform) that continuous functions on bounded domains can be approximated with arbitrary accuracy in the L^2 norm on that domain. Indeed the space of polynomials is dense in that Hilbert space.

Another important offshoot of approximation by arithmetic sums is that the Gibbs phenomenon doesn't occur. This follows easily from the next result.

Lemma 25 *Suppose the 2π-periodic function $f(t) \in L^2[-\pi, \pi]$ is bounded, with $M = \sup_{t \in [-\pi,\pi]} |f(t)|$. Then $|\sigma_n(t)| \le M$ for all t.*

PROOF: From (3.19) and (23) we have

$$|\sigma_k(t)| \le \frac{1}{2k\pi} \int_0^{2\pi} |f(t+x)| \left(\frac{\sin kx/2}{\sin x/2} \right)^2 dx \le \frac{M}{2k\pi} \int_0^{2\pi} \left(\frac{\sin kx/2}{\sin x/2} \right)^2 dx = M.$$

Q.E.D.

Now consider the example which has been our prototype for the Gibbs phenomenon:

$$h(t) = \begin{cases} 0, & t = 0 \\ \frac{\pi - t}{2}, & 0 < t < 2\pi \\ 0, & t = 2\pi. \end{cases}$$

and $h(t+2\pi) = h(t)$. Here the ordinary Fourier series gives

$$\frac{\pi - t}{2} = \sum_{n=1}^{\infty} \frac{\sin nt}{n}, \qquad 0 < t < 2\pi.$$

and this series exhibits the Gibbs phenomenon near the simple discontinuities at integer multiples of 2π. Furthermore the supremum of $|h(t)|$ is $\pi/2$ and it approaches the values $\pm\pi/2$ near the discontinuities. However, the lemma shows that $|\sigma(t)| < \pi/2$ for all n and t. Thus the arithmetic sums never overshoot or undershoot as t approaches the discontinuities. Thus there is no Gibbs phenomenon in the arithmetic series for this example.

In fact, the example is universal; there is no Gibbs phenomenon for arithmetic sums. To see this, we can mimic the proof of Theorem 23. This then shows that the arithmetic sums for all piecewise smooth functions converge uniformly except in arbitrarily small neighborhoods of the discontinuities of these functions. In the neighborhood of each discontinuity the arithmetic sums behave exactly as does the series for $h(t)$. Thus there is no overshooting or undershooting.

REMARK: The pointwise convergence criteria for the arithmetic means are much more general (and the proofs of the theorems are simpler) than for the case of ordinary Fourier series. Further, they provide a means of getting around the most serious problems caused by the Gibbs phenomenon. The technical reason for this is that the kernel function $F_k(t)$ is nonnegative. Why don't we drop ordinary Fourier series and just use the arithmetic means? There are a number of reasons, one being that the arithmetic means $\sigma_k(t)$are not the best L^2 approximations for order k, whereas the $S_k(t)$ *are* the best L^2 approximations. There is no Parseval theorem for arithmetic means. Further, once the approximation $S_k(t)$ is computed for ordinary Fourier series, in order to get the next level of approximation one needs only to compute two more constants:

$$S_{k+1}(t) = S_k(t) + a_{k+1}\cos(k+1)t + b_{k+1}\sin(k+1)t.$$

However, for the arithmetic means, in order to update $\sigma_k(t)$ to $\sigma_{k+1}(t)$ one must recompute *ALL* of the expansion coefficients. This is a serious practical difficulty.

Chapter 4

The Fourier Transform

4.1 The transform as a limit of Fourier series

We start by constructing the Fourier series (complex form) for functions on an interval $[-\pi L, \pi L]$. The ON basis functions are

$$e_n(t) = \frac{1}{\sqrt{2\pi L}} e^{\frac{int}{L}}, \qquad n = 0, \pm 1, \cdots,$$

and a sufficiently smooth function f of period $2\pi L$ can be expanded as

$$f(t) = \sum_{n=-\infty}^{\infty} \left(\frac{1}{2\pi L} \int_{-\pi L}^{\pi L} f(x) e^{-\frac{inx}{L}} dx \right) e^{\frac{int}{L}}.$$

For purposes of motivation let us abandon periodicity and think of the functions f as differentiable everywhere, vanishing at $t = \pm\pi L$ and identically zero outside $[-\pi L, \pi L]$. We rewrite this as

$$f(t) = \sum_{n=-\infty}^{\infty} e^{\frac{int}{L}} \frac{1}{2\pi L} \hat{f}(\frac{n}{L})$$

which looks like a Riemann sum approximation to the integral

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\lambda) e^{i\lambda t} d\lambda \tag{4.1}$$

to which it would converge as $L \to \infty$. (Indeed, we are partitioning the λ interval $[-L, L]$ into $2L$ subintervals, each with partition width $1/L$.) Here,

$$\hat{f}(\lambda) = \int_{-\infty}^{\infty} f(t) e^{-i\lambda t} dt. \tag{4.2}$$

Similarly the Parseval formula for f on $[-\pi L, \pi L]$,

$$\int_{-\pi L}^{\pi L} |f(t)|^2 dt = \sum_{n=-\infty}^{\infty} \frac{1}{2\pi L} |\hat{f}(\frac{n}{L})|^2$$

goes in the limit as $L \to \infty$ to the *Plancherel identity*

$$2\pi \int_{-\infty}^{\infty} |f(t)|^2 dt = \int_{-\infty}^{\infty} |\hat{f}(\lambda)|^2 d\lambda. \tag{4.3}$$

Expression (4.2) is called the *Fourier integral* or *Fourier transform* of f. Expression (4.1) is called the *inverse Fourier integral* for f. The Plancherel identity suggests that the Fourier transform is a one-to-one norm preserving map of the Hilbert space $L^2[-\infty, \infty]$ onto itself (or to another copy of itself). We shall show that this is the case. Furthermore we shall show that the pointwise convergence properties of the inverse Fourier transform are somewhat similar to those of the Fourier series. Although we could make a rigorous justification of the the steps in the Riemann sum approximation above, we will follow a different course and treat the convergence in the mean and pointwise convergence issues separately.

A second notation that we shall use is

$$\mathcal{F}[f](\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\lambda t} dt = \frac{1}{\sqrt{2\pi}} \hat{f}(\lambda) \tag{4.4}$$

$$\mathcal{F}^*[g](t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\lambda) e^{i\lambda t} d\lambda \tag{4.5}$$

Note that, formally, $\mathcal{F}^*[\hat{f}](t) = \sqrt{2\pi} f(t)$. The first notation is used more often in the engineering literature. The second notation makes clear that $\mathcal{F}$ and $\mathcal{F}^*$ are linear operators mapping $L^2[-\infty, \infty]$ onto itself in one view [and $\mathcal{F}$ mapping the *signal space* onto the *frequency space* with $\mathcal{F}^*$ mapping the frequency space onto the signal space in the other view. In this notation the Plancherel theorem takes the more symmetric form

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \int_{-\infty}^{\infty} |\mathcal{F}[f](\lambda)|^2 d\lambda.$$

EXAMPLES:

1. The box function (or rectangular wave)

$$\Pi(t) = \begin{cases} 1 & \text{if } -\pi < t < \pi \\ \frac{1}{2} & \text{if } t = \pm\pi \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

Then, since $\Pi(t)$ is an even function and $e^{-i\lambda t} = \cos(\lambda t) + i\sin(\lambda t)$, we have

$$\hat{\Pi}(\lambda) = \sqrt{2\pi}\mathcal{F}[\Pi](\lambda) = \int_{-\infty}^{\infty} \Pi(t)e^{-i\lambda t}dt = \int_{-\infty}^{\infty} \Pi(t)\cos(\lambda t)dt$$

$$= \int_{-\pi}^{\pi} \cos(\lambda t)dt = \frac{2\sin(\pi\lambda)}{\lambda} = 2\pi \text{ sinc } \lambda.$$

Thus sinc λ is the Fourier transform of the box function. The inverse Fourier transform is

$$\int_{-\infty}^{\infty} \text{sinc}(\lambda)e^{i\lambda t}d\lambda = \Pi(t),$$

as follows from (3.15). Furthermore, we have

$$\int_{-\infty}^{\infty} |\Pi(t)|^2 dt = 2\pi$$

and

$$\int_{-\infty}^{\infty} |\text{ sinc }(\lambda)|^2 d\lambda = 1$$

from (3.16), so the Plancherel equality is verified in this case. Note that the inverse Fourier transform converged to the midpoint of the discontinuity, just as for Fourier series.

2. A truncated cosine wave.

$$f(t) = \begin{cases} \cos 3t & \text{if } -\pi < t < \pi \\ -\frac{1}{2} & \text{if } t = \pm\pi \\ 0 & \text{otherwise.} \end{cases}$$

Then, since the cosine is an even function, we have

$$\hat{f}(\lambda) = \sqrt{2\pi}\mathcal{F}[f](\lambda) = \int_{-\infty}^{\infty} f(t)e^{i\lambda t}dt = \int_{-\pi}^{\pi} \cos(3t)\cos(\lambda t)dt$$

$$= \frac{2\lambda\sin(\lambda)}{9-\lambda^2}.$$

3. A truncated sine wave.

$$f(t) = \begin{cases} \sin 3t & \text{if } -\pi \le t \le \pi \\ 0 & \text{otherwise.} \end{cases}$$

Since the sine is an odd function, we have

$$\hat{f}(\lambda) = \sqrt{2\pi}\mathcal{F}[f](\lambda) = \int_{-\infty}^{\infty} f(t)e^{-i\lambda t}dt = -i\int_{-\pi}^{\pi} \sin(3t)\sin(\lambda t)dt$$

$$= \frac{-6i\sin(\lambda)}{9-4\lambda^2}.$$

4. A triangular wave.

$$f(t) = \begin{cases} \pi + t & \text{if } -\pi \le t \le 0 \\ \pi - t & \text{if } 0 \le t \le \pi \\ 0 & \text{otherwise}. \end{cases} \tag{4.7}$$

Then, since f is an even function, we have

$$\hat{f}(\lambda) = \sqrt{2\pi}\mathcal{F}[f](\lambda) = \int_{-\infty}^{\infty} f(t)e^{-i\lambda t}dt = 2\int_{0}^{\pi} (\pi - t)\cos(\lambda t)dt$$

$$= \frac{2-2\cos\lambda}{\lambda^2}.$$

NOTE: The Fourier transforms of the discontinuous functions above decay as $\frac{1}{\lambda}$ for $|\lambda| \to \infty$ whereas the Fourier transforms of the continuous functions decay as $\frac{1}{\lambda^2}$. The coefficients in the Fourier series of the analogous functions decay as $\frac{1}{n}$, $\frac{1}{n^2}$, respectively, as $|n| \to \infty$.

4.1.1 Properties of the Fourier transform

Recall that

$$\mathcal{F}[f](\lambda) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{-i\lambda t}dt = \frac{1}{\sqrt{2\pi}}\hat{f}(\lambda)$$

$$\mathcal{F}^*[g](t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(\lambda)e^{i\lambda t}d\lambda$$

We list some properties of the Fourier transform that will enable us to build a repertoire of transforms from a few basic examples. Suppose that f, g belong to $L^1[-\infty, \infty]$, i.e., $\int_{-\infty}^{\infty} |f(t)|dt < \infty$ with a similar statement for g. We can state the following (whose straightforward proofs are left to the reader):

1. $\mathcal{F}$ and $\mathcal{F}^*$ are linear operators. For $a, b \in C$ we have

$$\mathcal{F}[af + bg] = a\mathcal{F}[f] + b\mathcal{F}[g], \quad \mathcal{F}^*[af + bg] = a\mathcal{F}^*[f] + b\mathcal{F}^*[g].$$

2. Suppose $t^n f(t) \in L^1[-\infty, \infty]$ for some positive integer n. Then

$$\mathcal{F}[t^n f(t)](\lambda) = i^n \frac{d^n}{d\lambda^n}\{\mathcal{F}[f](\lambda)\}.$$

3. Suppose $\lambda^n f(\lambda) \in L^1[-\infty, \infty]$ for some positive integer n. Then

$$\mathcal{F}^*[\lambda^n f(\lambda)](t) = i^n \frac{d^n}{dt^n}\{\mathcal{F}^*[f](t)\}.$$

4. Suppose the nth derivative $f^{(n)}(t) \in L^1[-\infty, \infty]$ and piecewise continuous for some positive integer n, and f and the lower derivatives are all continuous in $(-\infty, \infty)$. Then

$$\mathcal{F}[f^{(n)}](\lambda) = (i\lambda)^n \mathcal{F}[f](\lambda)\}.$$

5. Suppose nth derivative $f^{(n)}(\lambda) \in L^1[-\infty, \infty]$ for some positive integer n and piecewise continuous for some positive integer n, and f and the lower derivatives are all continuous in $(-\infty, \infty)$. Then

$$\mathcal{F}^*[f^{(n)}](t) = (-it)^n \mathcal{F}^*[f](t).$$

6. The Fourier transform of a translation by real number a is given by

$$\mathcal{F}[f(t-a)](\lambda) = e^{-i\lambda a}\mathcal{F}[f](\lambda).$$

7. The Fourier transform of a scaling by positive number b is given by

$$\mathcal{F}[f(bt)](\lambda) = \frac{1}{b}\mathcal{F}[f](\frac{\lambda}{b}).$$

8. The Fourier transform of a translated and scaled function is given by

$$\mathcal{F}[f(bt-a)](\lambda) = \frac{1}{b}e^{-i\lambda a/b}\mathcal{F}[f](\frac{\lambda}{b}).$$

EXAMPLES

- We want to compute the Fourier transform of the rectangular box function with support on $[c, d]$:

$$R(t) = \begin{cases} 1 & \text{if } c < t < d \\ \frac{1}{2} & \text{if } t = c, d \\ 0 & \text{otherwise.} \end{cases}$$

  Recall that the box function

$$\Pi(t) = \begin{cases} 1 & \text{if } -\pi < t < \pi \\ \frac{1}{2} & \text{if } t = \pm\pi \\ 0 & \text{otherwise.} \end{cases}$$

  has the Fourier transform $\hat{\Pi}(\lambda) = 2\pi \operatorname{sinc} \lambda.$ but we can obtain R from Π by first translating $t \to s = t - \frac{(c+d)}{2}$ and then rescaling $s \to \frac{2\pi}{d-c}s$:

$$R(t) = \Pi(\frac{2\pi}{d-c}t - \pi\frac{c+d}{d-c}).$$

$$\hat{R}(\lambda) = \frac{4\pi^2}{d-c}e^{i\pi\lambda(c+d)/(d-c)}\text{sinc}(\frac{2\pi\lambda}{d-c}). \tag{4.8}$$

  Furthermore, from (3.15) we can check that the inverse Fourier transform of $\hat{R}$ is R, i.e., $\mathcal{F}^*(\mathcal{F})R(t) = R(t)$.

- Consider the truncated sine wave

$$f(t) = \begin{cases} \sin 3t & \text{if } -\pi \leq t \leq \pi \\ 0 & \text{otherwise} \end{cases}$$

  with

$$\hat{f}(\lambda) = \frac{-6i\sin(\lambda)}{9 - 4\lambda^2}.$$

  Note that the derivative f' of $f(t)$ is just $3g(t)$ (except at 2 points) where $g(t)$ is the truncated cosine wave

$$g(t) = \begin{cases} \cos 3t & \text{if } -\pi < t < \pi \\ -\frac{1}{2} & \text{if } t = \pm\pi \\ 0 & \text{otherwise.} \end{cases}$$

  We have computed

$$\hat{g}(\lambda) = \frac{2\lambda\sin(\lambda)}{9 - 4\lambda^2}.$$

  so $3\hat{g}(\lambda) = (i\lambda)\hat{f}(\lambda)$, as predicted.

- Reversing the example above we can differentiate the truncated cosine wave to get the truncated sine wave. The prediction for the Fourier transform doesn't work! Why not?

4.1.2 Fourier transform of a convolution

There is one property of the Fourier transform that is of particular importance in this course. Suppose f, g belong to $L^1[-\infty, \infty]$.

Definition 24 *The convolution of f and g is the function $f * g$ defined by*

$$(f * g)(t) = \int_{-\infty}^{\infty} f(t-x)g(x)dx.$$

Note also that $(f * g)(t) = \int_{-\infty}^{\infty} f(x)g(t-x)dx$, as can be shown by a change of variable.

Lemma 26 $f * g \in L^1[-\infty, \infty]$ *and*

$$\int_{-\infty}^{\infty} |f * g(t)|dt = \int_{-\infty}^{\infty} |f(x)|dx \int_{-\infty}^{\infty} |g(t)|dt.$$

SKETCH OF PROOF:

$$\int_{-\infty}^{\infty} |f * g(t)|dt = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |f(x)g(t-x)|dx \right) dt$$

$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |g(t-x)|dt \right) |f(x)|dx = \int_{-\infty}^{\infty} |g(t)|dt \int_{-\infty}^{\infty} |f(x)|dx.$$

Q.E.D.

Theorem 28 *Let $h = f * g$. Then*

$$\hat{h}(\lambda) = \hat{f}(\lambda)\hat{g}(\lambda).$$

SKETCH OF PROOF:

$$\hat{h}(\lambda) = \int_{-\infty}^{\infty} f * g(t)e^{-i\lambda t}dt = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x)g(t-x)dx \right) e^{-i\lambda t}dt$$

$$= \int_{-\infty}^{\infty} f(x)e^{-i\lambda x} \left(\int_{-\infty}^{\infty} g(t-x)e^{-i\lambda(t-x)}dt \right) dx = \int_{-\infty}^{\infty} f(x)e^{-i\lambda x}dx \ \hat{g}(\lambda)$$

$$= \hat{f}(\lambda)\hat{g}(\lambda).$$

Q.E.D.

4.2 L^2 convergence of the Fourier transform

In this course our primary interest is in Fourier transforms of functions in the Hilbert space $L^2[-\infty, \infty]$. However, the formal definition of the Fourier integral transform,

$$\mathcal{F}[f](\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\lambda t} dt \tag{4.9}$$

doesn't make sense for a general $f \in L^2[-\infty, \infty]$. If $f \in L^1[-\infty, \infty]$ then f is absolutely integrable and the integral (4.9) converges. However, there are square integrable functions that are not integrable. (Example: $f(t) = \frac{1}{1+|t|}$.) How do we define the transform for such functions?

We will proceed by defining $\mathcal{F}$ on a dense subspace of $f \in L^2[-\infty, \infty]$ where the integral makes sense and then take Cauchy sequences of functions in the subspace to define $\mathcal{F}$ on the closure. Since $\mathcal{F}$ preserves inner product, as we shall show, this simple procedure will be effective.

First some comments on integrals of L^2 functions. If $f, g \in L^2[-\infty, \infty]$ then the integral $(f, g) = \int_{-\infty}^{\infty} f(t)\overline{g}(t)dt$ necessarily exists, whereas the integral (4.9) may not, because the exponential $e^{-i\lambda t}$ is not an element of L^2. However, the integral of $f \in L^2$ over any finite interval, say $[-N, N]$ does exist. Indeed for N a positive integer, let $\chi_{[-N,N]}$ be the indicator function for that interval:

$$\chi_{[-N,N]}(t) = \begin{cases} 1 & \text{if } -N \leq t \leq N \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

Then $\chi_{[-N,N]} \in L^2[-\infty, \infty]$ so $\int_{-N}^{N} f(t)dt$ exists because

$$\int_{-N}^{N} |f(t)|dt = |(|f|, \chi_{[-N,N]})| \leq ||f||_{L^2}||\chi_{[-N,N]}||_{L^2} = ||f||_{L^2}\sqrt{2N} < \infty$$

Now the space of step functions is dense in $L^2[-\infty, \infty]$, so we can find a convergent sequence of step functions $\{s_n\}$ such that $\lim_{n\to\infty} ||f - s_n||_{L^2} = 0$. Note that the sequence of functions $\{f_N = f\chi_{[-N,N]}\}$ converges to f pointwise as $N \to \infty$ and each $f_N \in L^2 \cap L^1$.

Lemma 27 *$\{f_N\}$ is a Cauchy sequence in the norm of $L^2[-\infty, \infty]$ and $\lim_{n\to\infty} ||f - f_n||_{L^2} = 0$.*

PROOF: Given $\epsilon > 0$ there is step function s_M such that $||f - s_M|| < \frac{\epsilon}{2}$. Choose N so large that the support of s_M is contained in $[-N, N]$, i.e., $s_M(t)\chi_{[-N,N]}(t) =$

$s_M(t)$ for all t. Then $||s_M - f_N||^2 = \int_{-N}^{N} |s_M(t) - f(t)|^2 dt \leq \int_{-\infty}^{\infty} |s_M(t) - f(t)|^2 dt = ||s_M - f||^2$, so

$$||f - f_N|| - ||(f - s_M) + (s_M - f_N)|| \leq ||f - s_M|| + ||s_M - f_N|| \leq 2||f - s_M|| < \epsilon.$$

Q.E.D.

Here we will study the linear mapping $\mathcal{F} : L^2[-\infty, \infty] \to \hat{L}^2[-\infty, \infty]$ from the signal space to the frequency space. We will show that the mapping is *unitary*, i.e., it preserves the inner product and is 1-1 and onto. Moreover, the map $\mathcal{F}^* : \hat{L}^2[-\infty, \infty] \to L^2[-\infty, \infty]$ is also a unitary mapping and is the inverse of $\mathcal{F}$:

$$\mathcal{F}^*\mathcal{F} = I_{L^2}, \qquad \mathcal{F}\mathcal{F}^* = I_{\hat{L}^2}$$

where $I_{L^2}, I_{\hat{L}^2}$ are the identity operators on L^2 and $\hat{L}^2$, respectively. We know that the space of step functions is dense in L^2. Hence to show that $\mathcal{F}$ preserves inner product, it is enough to verify this fact for step functions and then go to the limit. Once we have done this, we can define $\mathcal{F}f$ for any $f \in L^2[-\infty, \infty]$. Indeed, if $\{s_n\}$ is a Cauchy sequence of step functions such that $\lim_{n\to\infty} ||f - s_n||_{L^2} = 0$, then $\{\mathcal{F}s_n\}$ is also a Cauchy sequence (indeed, $||s_n - s_m|| = ||\mathcal{F}s_n - \mathcal{F}s_m||$) so we can define $\mathcal{F}f$ by $\mathcal{F}f = \lim_{n\to\infty} \mathcal{F}s_n$. The standard methods of Section 1.3 show that $\mathcal{F}f$ is uniquely defined by this construction. Now the truncated functions f_N have Fourier transforms given by the convergent integrals

$$\mathcal{F}[f_N](\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-N}^{N} f(t) e^{-i\lambda t} dt$$

and $\lim_{N\to\infty} ||f - f_N||_{L^2} = 0$. Since $\mathcal{F}$ preserves inner product we have $||\mathcal{F}f - \mathcal{F}f_N||_{L^2} = ||\mathcal{F}(f - f_N)||_{L^2} = ||f - f_N||_{L^2}$, so $\lim_{N\to\infty} ||\mathcal{F}f - \mathcal{F}f_N||_{L^2} = 0$. We write

$$\mathcal{F}[f](\lambda) = \text{l.i.m.}_{N\to\infty}\mathcal{F}[f_N](\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-N}^{N} f(t) e^{-i\lambda t} dt$$

where 'l.i.m.' indicates that the convergence is in the mean (Hilbert space) sense, rather than pointwise.

We have already shown that the Fourier transform of the rectangular box function with support on $[c, d]$:

$$R_{c,d}(t) = \begin{cases} 1 & \text{if } c < t < d \\ \frac{1}{2} & \text{if } t = c, d \\ 0 & \text{otherwise.} \end{cases}$$

is

$$\mathcal{F}[R_{c,d}](\lambda) = \frac{4\pi^2}{\sqrt{2\pi}(d-c)} e^{i\pi\lambda(c+d)/(d-c)} \mathrm{sinc}(\frac{2\pi\lambda}{d-c}).$$

and that $\mathcal{F}^*(\mathcal{F})R_{c,d}(t) = R_{c,d}(t)$. (Since here we are concerned only with convergence in the mean the value of a step function at a particular point is immaterial. Hence for this discussion we can ignore such niceties as the values of step functions at the points of their jump discontinuities.)

Lemma 28

$$(R_{a,b}, R_{c,d})_{L^2} = (\mathcal{F}R_{a,b}, \mathcal{F}R_{c,d})_{\hat{L}^2}$$

for all real numbers $a \leq b$ and $c \leq d$.

PROOF:

$$(\mathcal{F}R_{a,b}, \mathcal{F}R_{c,d})_{\hat{L}^2} = \int_{-\infty}^{\infty} \mathcal{F}[R_{a,b}](\lambda)\overline{\mathcal{F}}[R_{c,d}](\lambda) d\lambda$$

$$= \lim_{N\to\infty} \int_{-N}^{N} \left(\mathcal{F}[R_{a,b}](\lambda) \int_c^d \frac{e^{i\lambda t}}{\sqrt{2\pi}} dt \right) d\lambda$$

$$= \lim_{N\to\infty} \int_c^d \left(\int_{-N}^{N} \mathcal{F}[R_{a,b}](\lambda) \frac{e^{i\lambda t}}{\sqrt{2\pi}} d\lambda \right) dt.$$

Now the inside integral is converging to $R_{a,b}$ as $N \to \infty$ in both the pointwise and L^2 sense, as we have shown. Thus

$$(\mathcal{F}R_{a,b}, \mathcal{F}R_{c,d})_{\hat{L}^2} = \int_c^d R_{a,b} dt = (R_{a,b}, R_{c,d})_{L^2}.$$

Q.E.D.

Since any step functions u, v are finite linear combination of indicator functions R_{a_j,b_j} with complex coefficients, $u = \sum_j \alpha_j R_{a_j,b_j}$, $v = \sum_k \beta_k R_{c_k,d_k}$ we have

$$(\mathcal{F}u, \mathcal{F}v)_{\hat{L}^2} = \sum_{j,k} \alpha_j \overline{\beta}_k (\mathcal{F}R_{a_j,b_j}, \mathcal{F}R_{c_k,d_k})_{\hat{L}^2}$$

$$= \sum_{j,k} \alpha_j \overline{\beta}_k (R_{a_j,b_j}, R_{c_k,d_k})_{L^2} = (u, v)_{L^2}.$$

Thus $\mathcal{F}$ preserves inner product on step functions, and by taking Cauchy sequences of step functions, we have the

Theorem 29 *(Plancherel Formula) Let $f, g \in L^2[-\infty, \infty]$. Then*

$$(f, g)_{L^2} = (\mathcal{F}f, \mathcal{F}g)_{\hat{L}^2}, \qquad ||f||^2_{L^2} = ||\mathcal{F}f||^2_{\hat{L}^2}$$

In the engineering notation this reads

$$2\pi \int_{-\infty}^{\infty} f(t)\overline{g}(t)dt = \int_{-\infty}^{\infty} \hat{f}(\lambda)\overline{\hat{g}}(\lambda)d\lambda.$$

Theorem 30 *The map $\mathcal{F}^* : \hat{L}^2[-\infty, \infty] \to L^2[-\infty, \infty]$ has the following properties:*

1. *It preserves inner product, i.e.,*

   $$(\mathcal{F}^*\hat{f}, \mathcal{F}^*\hat{g})_{L^2} = (\hat{f}, \hat{g})_{\hat{L}^2}$$

   for all $\hat{f}, \hat{g} \in \hat{L}^2[-\infty, \infty]$.

2. *$\mathcal{F}^*$ is the adjoint operator to $\mathcal{F} : L^2[-\infty, \infty] \to \hat{L}^2[-\infty, \infty]$, i.e.,*

   $$(\mathcal{F}f, \hat{g})_{\hat{L}^2} = (f, \mathcal{F}^*\hat{g})_{L^2},$$

   for all $f \in L^2[-\infty, \infty]$, $\hat{g} \in \hat{L}^2[-\infty, \infty]$.

PROOF:

1. This follows immediately from the facts that $\mathcal{F}$ preserves inner product and $\overline{\mathcal{F}}[\overline{f}](\lambda) = \mathcal{F}^*[f](\lambda)$.

2. $$(\mathcal{F}R_{a,b}, R_{c,d})_{\hat{L}^2} = (R_{a,b}, \mathcal{F}^*R_{c,d})_{L^2}$$

   as can be seen by an interchange in the order of integration. Then using the linearity of $\mathcal{F}$ and $\mathcal{F}^*$ we see that

   $$(\mathcal{F}u, v)_{\hat{L}^2} = (u, \mathcal{F}^*v)_{L^2},$$

   for all step functions u, v. Since the space of step functions is dense in $\hat{L}^2[-\infty, \infty]$ and in $L^2[-\infty, \infty]$

Q.E.D.

Theorem 31 1. *The Fourier transform $\mathcal{F} : L^2[-\infty, \infty] \to \hat{L}^2[-\infty, \infty]$ is a unitary transformation, i.e., it preserves the inner product and is 1-1 and onto.*

2. *The adjoint map $\mathcal{F}^* : \hat{L}^2[-\infty, \infty] \to L^2[-\infty, \infty]$ is also a unitary mapping.*

3. *$\mathcal{F}^*$ is the inverse operator to $\mathcal{F}$:*

$$\mathcal{F}^*\mathcal{F} = I_{L^2}, \qquad \mathcal{F}\mathcal{F}^* = I_{\hat{L}^2}$$

where $I_{L^2}, I_{\hat{L}^2}$ are the identity operators on L^2 and $\hat{L}^2$, respectively.

PROOF:

1. The only thing left to prove is that for every $\hat{g} \in \hat{L}^2[-\infty, \infty]$ there is a $f \in L^2[-\infty, \infty]$ such that $\mathcal{F}f = \hat{g}$, i.e., $\mathcal{R} \equiv \{\mathcal{F}f : f \in L^2[-\infty, \infty]\} = \hat{L}^2[-\infty, \infty]$. Suppose this isn't true. Then there exists a nonzero $\hat{h} \in \hat{L}^2[-\infty, \infty]$ such that $\hat{h} \perp \mathcal{R}$, i.e., $(\mathcal{F}f, \hat{h})_{\hat{L}^2} = 0$ for all $f \in L^2[-\infty, \infty]$. But this means that $(f, \mathcal{F}^*\hat{h})_{L^2} = 0$ for all $f \in L^2[-\infty, \infty]$, so $\mathcal{F}^*\hat{h} = \Theta$. But then $||\mathcal{F}^*\hat{h}||_{L^2} = ||\hat{h}||_{\hat{L}^2} = 0$ so $\hat{h} = \Theta$, a contradiction.

2. Same proof as for 1.

3. We have shown that $\mathcal{F}\mathcal{F}^*R_{a,b} = \mathcal{F}^*\mathcal{F}R_{a,b} = R_{a,b}$ for all indicator functions $R_{a,b}$. By linearity we have $\mathcal{F}\mathcal{F}^*s = \mathcal{F}^*\mathcal{F}s = s$for all step functions s. This implies that
$$(\mathcal{F}^*\mathcal{F}f, g)_{L^2} = (f, g)_{L^2}$$
for all $f, g \in L^2[-\infty, \infty]$. Thus
$$([\mathcal{F}^*\mathcal{F} - I_{L^2}]f, g)_{L^2} = 0$$
for all $f, g \in L^2[-\infty, \infty]$. Thus $\mathcal{F}^*\mathcal{F} = I_{L^2}$. An analogous argument gives $\mathcal{F}\mathcal{F}^* = I_{\hat{L}^2}$.

Q.E.D.

4.3 The Riemann-Lebesgue Lemma and pointwise convergence

Lemma 29 *(Riemann-Lebesgue) Suppose f is absolutely Riemann integrable in $(-\infty, \infty)$ (so that $f \in L^1[-\infty, \infty]$), and is bounded in any finite subinterval $[a, b]$, and let α, β be real. Then*

$$\lim_{\alpha \to +\infty} \int_{-\infty}^{\infty} f(t) \sin(\alpha t + \beta) dt = 0.$$

PROOF: Without loss of generality, we can assume that f is real, because we can break up the complex integral into its real and imaginary parts.

1. The statement is true if $f = R_{a,b}$ is an indicator function, for

   $$\int_{-\infty}^{\infty} R_{a,b}(t) \sin(\alpha t + \beta) dt = \int_a^b \sin(\alpha t + \beta) dt = \frac{-1}{\alpha} \cos(\alpha t + \beta)|_a^b \to 0$$

   as $\alpha \to +\infty$.

2. The statement is true if f is a step function, since a step function is a finite linear combination of indicator functions.

3. The statement is true if f is bounded and Riemann integrable on the finite interval $[a, b]$ and vanishes outside the interval. Indeed given any $\epsilon > 0$ there exist two step functions $\overline{s}$ (Darboux upper sum) and $\underline{s}$ (Darboux lower sum) with support in $[a, b]$ such that $\overline{s}(t) \geq f(t) \geq \underline{s}(t)$ for all $t \in [a, b]$ and $\int_a^b |\overline{s} - \underline{s}| < \frac{\epsilon}{2}$. Then

   $$\int_a^b f(t) \sin(\alpha t + \beta) dt = \int_a^b [f(t) - \underline{s}(t)] \sin(\alpha t + \beta) dt + \int_a^b \underline{s}(t) \sin(\alpha t + \beta) dt.$$

   Now

   $$|\int_a^b [f(t) - \underline{s}(t)] \sin(\alpha t + \beta) dt| \leq \int_a^b |f(t) - \underline{s}(t)| dt \leq \int_a^b |\overline{s} - \underline{s}| < \frac{\epsilon}{2}$$

   and (since $\underline{s}$ is a step function, by choosing α sufficiently large we can ensure

   $$|\int_a^b \underline{s}(t) \sin(\alpha t + \beta) dt| < \frac{\epsilon}{2}.$$

   Hence

   $$|\int_a^b f(t) \sin(\alpha t + \beta) dt| < \epsilon$$

   for α sufficiently large.

4. The statement of the lemma is true in general. Indeed

$$|\int_{-\infty}^{\infty} f(t)\sin(\alpha t+\beta)dt| \leq |\int_{-\infty}^{a} f(t)\sin(\alpha t+\beta)dt|$$

$$+|\int_{a}^{b} f(t)\sin(\alpha t+\beta)dt| + |\int_{b}^{\infty} f(t)\sin(\alpha t+\beta)dt|.$$

Given $\epsilon > 0$ we can choose a and b such the first and third integrals are each $< \frac{\epsilon}{3}$, and we can choose α so large the the second integral is $< \frac{\epsilon}{3}$. Hence the limit exists and is 0.

Q.E.D.

Theorem 32 *Let f be a complex valued function such that*

- *$f(t)$ is absolutely Riemann integrable on $(-\infty,\infty)$ (hence $f \in L^1[-\infty,\infty]$).*
- *$f(t)$ is piecewise continuous on $(-\infty,\infty)$, with only a finite number of discontinuities in any bounded interval.*
- *$f'(t)$ is piecewise continuous on $(-\infty,\infty)$, with only a finite number of discontinuities in any bounded interval.*
- *$f(t) = \frac{f(t+0)+f(t-0)}{2}$ at each point t.*

Let

$$\hat{f}(\lambda) = \int_{-\infty}^{\infty} f(t)e^{-i\lambda t}dt$$

be the Fourier transform of f. Then

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\lambda)e^{i\lambda t}d\lambda$$

for every $t \in (-\infty,\infty)$.

PROOF: For real $L > 0$ set

$$f_L(t) = \int_{-L}^{L} \hat{f}(\lambda)e^{i\lambda t}d\lambda = \frac{1}{2\pi}\int_{-L}^{L}\left[\int_{-\infty}^{\infty} f(x)e^{-i\lambda x}dx\right]e^{i\lambda t}d\lambda$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)\left[\int_{-L}^{L} e^{i\lambda(t-x)}d\lambda\right]dx = \int_{-\infty}^{\infty} f(x)\Delta_L(t-x)dx,$$

where

$$\Delta_L(x) = \frac{1}{2\pi}\int_{-L}^{L} e^{i\lambda x} d\lambda = \begin{cases} \frac{L}{\pi} & \text{if } x = 0 \\ \frac{\sin Lx}{\pi x} & \text{otherwise.} \end{cases}$$

Using the integral (3.13) we have,

$$f_L(t) - f(t) = \int_{-\infty}^{\infty} \Delta_L(t-x)[f(x) - f(t)]dx$$

$$\int_0^{\infty} \Delta_L(x)[f(t+x) + f(t-x) - 2f(t)]dx$$

$$= \int_0^{\infty} \{\frac{f(t+x) + f(t-x) - 2f(t)}{\pi x}\} \sin Lx \, dx$$

The function in the curly braces satisfies the assumptions of the Riemann-Lebesgue Lemma. Hence $\lim_{L\to+\infty}[f_L(t) - f(t)] = 0$. Q.E.D

Note: Condition 4 is just for convenience; redefining f at the discrete points where there is a jump discontinuity doesn't change the value of any of the integrals. The inverse Fourier transform converges to the midpoint of a jump discontinuity, just as does the Fourier series.

4.4 Relations between Fourier series and Fourier integrals: sampling, periodization

Definition 25 *A function f is said to be frequency band-limited if there exists a constant $\Omega > 0$ such that $\hat{f}(\lambda) = 0$ for $|\lambda| > \Omega$. The frequency $\nu = \frac{\Omega}{2\pi}$ is called the Nyquist frequency and 2ν is the Nyquist rate.*

Theorem 33 *(Shannon-Whittaker Sampling Theorem) Suppose f is a function such that*

1. *f satisfies the hypotheses of the Fourier convergence theorem 32.*

2. *$\hat{f}$ is continuous and has a piecewise continuous first derivative on its domain.*

3. *There is a fixed $\Omega > 0$ such that $\hat{f}(\lambda) = 0$ for $|\lambda| > \Omega$.*

Then f is completely determined by its values at the points $t_j = \frac{j\pi}{\Omega}, j = 0, \pm 1, \pm 2, \cdots$:

$$f(t) = \sum_{-\infty}^{\infty} f(\frac{j\pi}{\Omega}) \frac{\sin(\Omega t - j\pi)}{\Omega t - j\pi},$$

and the series converges uniformly on $(-\infty, \infty)$.

(NOTE: The theorem states that for a frequency band-limited function, to determine the value of the function at all points, it is sufficient to sample the function at the Nyquist rate, i.e., at intervals of $\frac{\pi}{\Omega}$. The method of proof is obvious: compute the Fourier series expansion of $\hat{f}(\lambda)$ on the interval $[-\Omega, \Omega]$.) PROOF: We have

$$\hat{f}(\lambda) = \sum_{k=-\infty}^{\infty} c_k e^{\frac{i\pi k\lambda}{\Omega}}, \qquad c_k = \frac{1}{2\Omega} \int_{-\Omega}^{\Omega} \hat{f}(\lambda) e^{-\frac{i\pi k\lambda}{\Omega}} d\lambda,$$

where the convergence is uniform on $[-\Omega, \Omega]$. This expansion holds only on the interval; $\hat{f}(\lambda)$ vanishes outside the interval.

Taking the inverse Fourier transform we have

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\lambda) e^{i\lambda t} d\lambda = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \hat{f}(\lambda) e^{i\lambda t} d\lambda$$

$$= \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \sum_{k=-\infty}^{\infty} c_k e^{\frac{i(\pi k + t\Omega)\lambda}{\Omega}} d\lambda$$

$$= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} c_k \int_{-\Omega}^{\Omega} e^{\frac{i(\pi k + t\Omega)\lambda}{\Omega}} d\lambda = \sum_{k=-\infty}^{\infty} c_k \frac{\Omega \sin(\Omega t + k\pi)}{\pi(\Omega t + k\pi)}.$$

Now

$$c_k = \frac{1}{2\Omega} \int_{-\Omega}^{\Omega} \hat{f}(\lambda) e^{-\frac{i\pi k\lambda}{\Omega}} d\lambda = \frac{1}{2\Omega} \int_{-\infty}^{\infty} \hat{f}(\lambda) e^{-\frac{i\pi k\lambda}{\Omega}} d\lambda = \frac{\pi}{\Omega} f(-\frac{\pi k}{\Omega}).$$

Hence, setting $k = -j$,

$$f(t) = \sum_{j=-\infty}^{\infty} f(\frac{j\pi}{\Omega}) \frac{\sin(\Omega t - j\pi)}{\Omega t - j\pi}.$$

Q.E.D.

Note: There is a trade-off in the choice of Ω. Choosing it as small as possible reduces the sampling rate. However, if we increase the sampling rate, i.e. *oversample*, the series converges more rapidly.

Another way to compare the Fourier transform $(-\infty,\infty)$ with Fourier series is to periodize a function. To get convergence we need to restrict ourselves to functions that decay rapidly at infinity. We could consider functions with compact support, say infinitely differentiable. Another useful but larger space of functions is the Schwartz class. We say that $f \in L^2[-\infty,\infty]$ belongs to the *Schwartz class* if f is infinitely differentiable everywhere, and there exist constants $C_{n,q}$ (depending on f) such that $|t^n \frac{d^q}{dt^q} f| \leq C_{n,q}$ on R for each $n, q = 0, 1, 2, \cdots$. Then the projection operator P maps an f in the Schwartz class to a continuous function in $L^2[0, 2\pi]$ with period 2π. (However, periodization can be applied to a much larger class of functions, e.g. functions on $L^2[-\infty,\infty]$ that decay as $\frac{c}{t^2}$ as $|t| \to \infty$.):

$$P[f](t) = \sum_{m=-\infty}^{\infty} f(t + 2\pi m) \qquad (4.11)$$

Expanding $P[f](t)$ into a Fourier series we find

$$P[f](t) = \sum_{n=-\infty}^{\infty} c_n e^{int}$$

where

$$c_n = \frac{1}{2\pi}\int_0^{2\pi} P[f](t)e^{-int}dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{-int}dx = \frac{1}{2\pi}\hat{f}(n)$$

where $\hat{f}(\lambda)$ is the Fourier transform of $f(t)$. Thus,

$$\sum_{n=-\infty}^{\infty} f(t + 2\pi n) = \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} \hat{f}(n)e^{int}, \qquad (4.12)$$

and we see that $P[f](t)$ tells us the value of $\hat{f}$ at the integer points $\lambda = n$, but not in general at the non-integer points. (For $t = 0$, equation (4.12) is known as the *Poisson summation formula*. If we think of f as a signal, we see that **periodization** (4.11) of f results in a loss of information. However, if f vanishes outside of $[0, 2\pi)$) then $P[f](t) \equiv f(t)$ for $0 \leq t < 2\pi$ and

$$f(t) = \sum_n \hat{f}(n)e^{int}, \quad 0 \leq t < 2\pi$$

without error.)

4.5 The Fourier integral and the uncertainty relation of quantum mechanics

The uncertainty principle gives a limit to the degree that a function $f(t)$ can be simultaneously localized in time as well as in frequency. To be precise, we introduce some notation from probability theory. Every $f \in L^2[-\infty, \infty]$ defines a probability distribution function $\rho(t) = \frac{|f(t)|^2}{||f||^2}$, i.e., $\rho(t) \geq 0$ and $\int_{-\infty}^{\infty} \rho(t)dt = 1$.

Definition 26 • *The mean of the distribution defined by f is*

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

- *The dispersion of f about $a \in R$ is*

$$\Delta_a f = \frac{\int_{-\infty}^{\infty}(t-a)^2|f(t)|^2 dt}{\int_{-\infty}^{\infty} |f(t)|^2 dt}.$$

($\Delta_{\bar{t}} f$ is called the variance of f, and $\sqrt{\Delta_{\bar{t}} f}$ the standard deviation.

The dispersion of f about a is a measure of the extent to which the graph of f is concentrated at a. If $f = \delta(x-a)$ the “Dirac delta function”, the dispersion is zero. The constant $f(t) \equiv 1$ has infinite dispersion. (However there are no such L^2 functions.) Similarly we can define the dispersion of the Fourier transform of f about some point $\alpha \in R$:

$$\Delta_\alpha \hat{f} = \frac{\int_{-\infty}^{\infty}(\lambda-\alpha)^2|\hat{f}(\lambda)|^2 d\lambda}{\int_{-\infty}^{\infty} |\hat{f}(\lambda)|^2 d\lambda}.$$

Note: It makes no difference which definition of the Fourier transform that we use, $\hat{f}$ or $\mathcal{F}f$, because the normalization gives the same probability measure.

Example 3 *Let $f_s(t) = (\frac{2s}{\pi})^{1/4} e^{-st^2}$ for $s > 0$, the Gaussian distribution. From the fact that $\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}$ we see that $||f_s|| = 1$. The normed Fourier transform of f_s is $\hat{f}_s(\lambda) = (\frac{2}{s\pi})^{1/4} e^{\frac{-\lambda^2}{4s}}$. By plotting some graphs one can see informally that as s increases the graph of f_s concentrates more and more about $t = 0$, i.e.,* *the dispersion $\Delta_0 f_s$ decreases. However, the dispersion of $\hat{f}_s$ increases as s increases. We can't make both values, simultaneously, as small as we would like. Indeed, a straightforward computation gives*

$$\Delta_0 f_s = \frac{1}{4s}, \qquad \Delta_0 \hat{f}_s = s,$$

so the product of the variances of f_s and $\hat{f}_s$ is always $\frac{1}{4}$, no matter how we choose s.

Theorem 34 *(Heisenberg inequality, Uncertainty theorem) If $f(t) \neq 0$ and $tf(t)$ belong to $L^2[-\infty,\infty]$ then $\Delta_a f \Delta_\alpha \hat{f} \geq \frac{1}{4}$ for any $a, \alpha \in R$.*

SKETCH OF PROOF: I will give the proof under the added assumptions that $f'(t)$ exists everywhere and also belongs to $L^2[-\infty,\infty]$. (In particular this implies that $f(t) \to 0$ as $t \to \pm\infty$.) The main ideas occur there.

We make use of the canonical commutation relation of quantum mechanics, the fact that the operations of multiplying a function $f(t)$ by t, $(Tf(t) = tf(t))$ and of differentiating a function $(Df(t) = f'(t))$ don't commute: $DT - TD = I$. Thus

$$\frac{d}{dt}[tf(t)] - t\left[\frac{d}{dt}f(t)\right] = f(t).$$

Now it is easy from this to check that

$$(\frac{d}{dt} - i\alpha)\left[(t-a)f(t)\right] - (t-a)\left[(\frac{d}{dt} - i\alpha)f(t)\right] = f(t)$$

also holds, for any $a, \alpha \in R$. (The a, α dependence just cancels out.) This implies that

$$\left((\frac{d}{dt} - i\alpha)[(t-a)f(t)], f(t)\right) - \left((t-a)[(\frac{d}{dt} - i\alpha)f(t)], f(t)\right) = (f(t), f(t)) = ||f||^2.$$

Integrating by parts in the first integral, we can rewrite the identity as

$$-\left([(t-a)f(t)], [(\frac{d}{dt} - i\alpha)f(t)]\right) - \left([(\frac{d}{dt} - i\alpha)f(t)], [(t-a)f(t)]\right) = ||f||^2.$$

The Schwarz inequality and the triangle inequality now yield

$$||f||^2 \leq 2||(t-a)f(t)|| \cdot ||(\frac{d}{dt} - i\alpha)f(t)||. \tag{4.13}$$

¿From the list of properties of the Fourier transform in Section 4.1.1 and the Plancherel formula, we see that $||(\frac{d}{dt} - i\alpha)f(t)|| = \frac{1}{\sqrt{2\pi}}||(\lambda - \alpha)\hat{f}(\lambda)||$ and $||f|| = \frac{1}{\sqrt{2\pi}}||\hat{f}||$. Then, dividing by $||f||$ and squaring, we have

$$\Delta_a f \Delta_\alpha \hat{f} \geq \frac{1}{4}.$$

Q.E.D.

NOTE: Normalizing to $a = \alpha = 0$ we see that the Schwarz inequality becomes an equality if and only if $2stf(t) + \frac{d}{dt}f(t) = 0$ for some constant s. Solving this differential equation we find $f(t) = c_0 e^{-st^2}$ where c_0 is the integration constant, and we must have $s > 0$ in order for f to be square integrable. Thus the Heisenberg inequality becomes an equality only for Gaussian distributions.

Chapter 5

Discrete Fourier Transform

5.1 Relation to Fourier series: aliasing

Suppose that f is a square integrable function on the interval $[0, 2\pi]$, periodic with period 2π, and that the Fourier series expansion converges pointwise to f everywhere:

$$f(t) = \sum_n \hat{f}(n) e^{int}, \quad 0 \le t \le 2\pi \tag{5.1}$$

What is the effect of sampling the signal at a finite number of equally spaced points? For an integer $N > 1$ we sample the signal at $2\pi m/N, m = 0, 1, \cdots, N-1$:

$$f\left(\frac{2\pi m}{N}\right) = \sum_n \hat{f}(n) e^{2\pi i n m/N}, \quad 0 \le m < N.$$

¿From the Euclidean algorithm we have $n = a + bN$ where $0 \le a < N$ and a, b are integers. Thus

$$f\left(\frac{2\pi m}{N}\right) = \sum_{a=0}^{N-1} \left[\sum_b \hat{f}(a + bN)\right] e^{2\pi i m a/N}, \quad 0 \le m < N. \tag{5.2}$$

Note that the quantity in brackets is the projection of $\hat{f}$ at integer points to a periodic function of period N. Furthermore, the expansion (5.2) is essentially the *finite Fourier expansion*, as we shall see. However, simply sampling the signal at the points $2\pi m/N$ tells us only $\sum_b \hat{f}(a + bN)$, not (in general) $\hat{f}(a)$. This is known as **aliasing error**. If f is sufficiently smooth and N sufficiently large that all of the Fourier coefficients $\hat{f}(n)$ for $n > N$ can be neglected, then this gives a good approximation of the Fourier series.

5.2 The definition

To further motivate the Discrete Fourier Transform (DFT) it is useful to consider the periodic function $f(t)$ above as a function on the unit circle: $f(t) = g(e^{it})$. Thus t corresponds to the point $(x, y) = (\cos t, \sin t)$ on the unit circle, and the points with coordinates t and $t + 2\pi n$ are identified, for any integer n. In the complex plane the points on the unit circle would just be e^{it}. Given an integer $N > 0$, let us sample f at N points $e^{\frac{2\pi i}{N}n}$, $n = 0, 1, \cdots, N-1$, evenly spaced around the unit circle. We denote the value of f at the nth point by f[n], and the full set of values by the column vector

$$f = (f[0], f[1], \cdots f[N-1]). \tag{5.3}$$

It is useful to extend the definition of $f[n]$ for all integers n by the periodicity requirement $f[n] = f[n + kN]$ for all integers k, i.e., $f[n] = f[m]$ if $n = m \bmod N$. (This is precisely what we should do if we consider these values to be samples of a function on the unit circle.)

We will consider the vectors (5.3) as belonging to an N-dimensional inner product space P_N and expand f in terms of a specially chosen ON basis. To get the basis functions we sample the Fourier basis functions $e_k(t) = e^{ikt}$ around the unit circle:

$$e_k[n] = e^{\frac{2\pi i}{N}nk} = \omega^{-nk}$$

or as a column vector

$$e_k = (e_k[0], e_k[1], \cdots e_k[N-1]) = (1, \omega^{-k}, \omega^{-2k}, \cdots \omega^{-(N-1)k}), \tag{5.4}$$

where ω is the primitive Nth root of unity $\omega = e^{-\frac{2\pi i}{N}}$.

Lemma 30

$$\sum_{n=0}^{N-1} \omega^{kn} = \begin{cases} 0 & \mathit{if}\, k = 1, 2, \cdots, N-1, \bmod N \\ N & \qquad \mathit{if}\, k = 0, \bmod N \end{cases}$$

PROOF: Since $\omega^N = 1$ and $\omega \neq 1$ we have

$$1 - \omega^N = 0 = (1 - \omega)(1 + \omega + \omega^2 + \cdots + \omega^{N-1}).$$

Thus

$$\sum_{n=0}^{N-1} \omega^n = 1 + \omega + \omega^2 + \cdots + \omega^{N-1} = 0.$$

Since ω^k is also an Nth root of unity and $\omega^k \neq 1$ for $k = 1, \cdots, N-1$, so the same argument shows that $\sum_{n=0}^{N-1} \omega^{kn} = 0$. However, if k=0 the sum is N. Q.E.D.

We define an inner product on P_N by

$$(f, g)_N = \frac{1}{N} \sum_{n=0}^{N-1} f[n]\overline{g}[n], \qquad f[n], g[n] \in P_N.$$

Lemma 31 *The functions e_k, $k = 0, 1, \cdots, N-1$ form an ON basis for P_N.*

PROOF:

$$(e_j, e_k)_N = \frac{1}{N} \sum_{n=0}^{N-1} e_j[n]\overline{e}_k[n] = \frac{1}{N} \sum_{n=0}^{N-1} \omega^{(k-j)n} = \begin{cases} 0 & \text{if } j \neq k \bmod N \\ 1 & \text{if } j = k \bmod N. \end{cases}$$

Thus $(e_j, e_k)_N = \delta_{jk}$ where the result is understood mod N. Now we can expand $f \sim \{f[n]\}$ in terms of this ON basis:

$$f = \frac{1}{N} \sum_{k=0}^{N-1} F[k] e_k,$$

or in terms of components,

$$f[n] = \frac{1}{N} \sum_{k=0}^{N-1} F[k] e^{2\pi i k n/N} = \frac{1}{N} \sum_{k=0}^{N-1} F[k]\omega^{-nk}. \tag{5.5}$$

The Fourier coefficients $F[k]$ of f are computed in the standard way: $F[k]/N = (f, e_k)_N$ or

$$F[k] = \sum_{n=0}^{N-1} f[n] e_k[-n] = \sum_{n=0}^{N-1} f[n]\omega^{kn}. \tag{5.6}$$

The Parseval (Plancherel) equality reads

$$\sum_{n=0}^{N-1} f[n]\overline{g}[n] = \sum_{k=0}^{N-1} F[k]\overline{G}[k]$$

for $f, g \in P_N$.

The column vector

$$F = (F[0], F[1], F[2], \cdots, F[N-1])$$

is the *Discrete Fourier transform* (DFT) of $f = \{f[n]\}$. It is illuminating to express the discrete Fourier transform and its inverse in matrix notation. The DFT is given by the matrix equation $F = \mathcal{F}_N f$ or

$$\begin{pmatrix} F[0] \\ F[1] \\ F[2] \\ \vdots \\ F[N-1] \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{pmatrix} \begin{pmatrix} f[0] \\ f[1] \\ f[2] \\ \vdots \\ f[N-1] \end{pmatrix}. \tag{5.7}$$

Here $\mathcal{F}_N$ is an $N \times N$ matrix. The inverse relation is the matrix equation $f = \mathcal{F}_N^{-1} F$ or

$$\begin{pmatrix} f[0] \\ f[1] \\ f[2] \\ \vdots \\ f[N-1] \end{pmatrix} = \frac{1}{N} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \bar{\omega} & \bar{\omega}^2 & \cdots & \bar{\omega}^{N-1} \\ 1 & \bar{\omega}^2 & \bar{\omega}^4 & \cdots & \bar{\omega}^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{\omega}^{N-1} & \bar{\omega}^{2(N-1)} & \cdots & \bar{\omega}^{(N-1)(N-1)} \end{pmatrix} \begin{pmatrix} F[0] \\ F[1] \\ F[2] \\ \vdots \\ F[N-1] \end{pmatrix}, \tag{5.8}$$

where $\omega = e^{-2\pi i/N}, \bar{\omega} = \omega^{-1} = e^{2\pi i/N}$.

NOTE: At this point we can drop any connection with the sampling of values of a function on the unit circle. The DFT provides us with a method of analyzing any N-tuple of values f in terms of Fourier components. However, the association with functions on the unit circle is a good guide to our intuition concerning when the DFT is an appropriate tool

Examples 4

1. $$f[n] = \delta[n] = \begin{cases} 1 & \textit{if } n = 0 \\ 0 & \textit{otherwise}. \end{cases}$$

   Here, $F[k] = 1$.

2. $f[n] = 1$ *for all* n. *Then* $F[k] = N\delta[k]$.

3. $f[n] = r^n$ *for* $n = 0, 1, \cdots, N-1$ *and* $r \in C$. *here*

$$F[k] = \begin{cases} N & \textit{if } r = e^{2\pi ik/N} \\ \frac{r^N - 1}{(re^{-2\pi ik/N} - 1)} & \textit{otherwise} \end{cases}$$

4. *Upsampling. Given $f \in P_{N'}$ where $N' = \frac{N}{2}$ we define $g \in P_N$ by*

$$g[n] = \begin{cases} f[\frac{n}{2}] & \textit{if } n = 0, \pm 2, \pm 4, \cdots \\ 0 & \textit{otherwise}. \end{cases}$$

   Then $G[k] = F[k]$ where $F[k]$ is periodic with period $N/2$.

5. *Downsampling. Given $f \in P_{2N}$ we define $g \in P_N$ by $g[n] = f[2n]$, $n = 0, 1, \cdots, N$. Then $G[k] = \frac{1}{2}(F[k] + F[k+N])$.*

5.2.1 More properties of the DFT

Note that if $f[n]$ is defined for all integers n by the periodicity property, $f[n + jN] = f[n]$ for all integers j, the transform $F[k]$ has the same property. Indeed $F[k] = \sum_{n=0}^{N-1} f[n]\omega^{kn}$, so $F[k + jN] = \sum_{n=0}^{N-1} f[n]\omega^{(k+jN)n} = F[k]$, since $\omega^N = 1$.

Here are some other properties of the DFT. Most are just observations about what we have already shown. A few take some proof.

Lemma 32

- *Symmetry. $\mathcal{F}_N^{tr} = \mathcal{F}_N$*

- *Unitarity. $\mathcal{F}_N^{-1} = \frac{1}{N}\overline{\mathcal{F}}_N$*

  Set $\mathcal{G}_N = \frac{1}{\sqrt{N}}\mathcal{F}_N$. Then $\mathcal{G}_N$ is unitary. That is $\mathcal{G}_N^{-1} = \overline{\mathcal{G}_N}^{tr}$. Thus the row vectors of $\mathcal{G}_N$ are mutually orthogonal and of length 1.

- *Let $S : P_N \to P_N$ be the shift operator on P_N. That is $Sf[n] = f[n-1]$ for any integer n. Then $SF[k] = \omega^k F[k]$ for any integer k. Further, $S^{-1}f[n] = f[n+1]$ and $S^{-1}F[k] = \omega^{-k}F[k]$.*

- *Let $M : P_N \to P_N$ be an operator on P_N such that $Mf[n] = \omega^{-n}f[n]$ for any integer n. Then $MF[k] = F[k+1]$ for any integer k.*

- *If $f = \{f[n]\}$ is a real vector then $F[N-k] = \overline{F}[k]$.*

- *For $f, g \in P_N$ define the convolution $f * g \in P_N$ by*

$$f * g[n] = \sum_{m=0}^{N-1} f[m]g[n-m].$$

  Then $f * g[n] = g * f[n]$ and $f * g[n + jN] = f * g[n]$.

- *Let* $h[n] = f * g[n]$*. Then* $H[k] = F[k]G[k]$*.*

Here are some simple examples of basic transformations applied to the 4-vector $(f[0], f[1], f[2], f[3])$ with DFT $(F[0], F[1], F[2], F[3])$ and $\omega = e^{-i\pi/2}$:

Operation	Data vector	DFT
Left shift	$(f[1], f[2], f[3], f[0])$	$(F[0], \omega^{-1}F[1], \omega^{-2}F[2], \omega^{-3}F[3])$
Right shift	$(f[3], f[0], f[1], f[2])$	$(F[0], \omega^{1}F[1], \omega^{2}F[2], \omega^{3}F[3])$
Upsampling	$(f[0], 0, f[1], 0, f[2], 0, f[3], 0)$	$(F[0], F[1], F[2], F[3], F[0], F[1], F[2], F[3])$
Downsampling	$(f[0], f[2])$	$\frac{1}{2}(F[0] + F[2], F[1] + F[3])$

(5.9)

5.2.2 An application of the DFT to finding the roots of polynomials

It is somewhat of a surprise that there is a simple application of the DFT to find the Cardan formulas for the roots r_0, r_1, r_2 of a third order polynomial:

$$P_3(x) = x^3 + ax^2 + bx + c = (x - r_0)(x - r_1)(x - r_2). \qquad (5.10)$$

Let ω be a primitive cube root of unity and define the 3-vector $(F[0], F[1], F[2]) = (r_0, r_1, r_2)$. Then

$$F[k] = r_k = f[0] + f[1]\omega^k + f[2]\omega^{2k}, \quad k = 0, 1, 2 \qquad (5.11)$$

where

$$f[n] = z_n = \frac{1}{3}\left(F[0] + F[1]\omega^{-n} + F[2]\omega^{-2n}\right), \quad n = 0, 1, 2.$$

Substituting relations (5.11) for the r_k into (5.10), we can expand the resulting expression in terms of the transforms z_n and powers of ω (remembering that $\omega^3 = 1$):

$$P_3(x) = A(x, z_n) + B(x, z_n)\omega + C(x, z_n)\omega^2.$$

This expression would appear to involve powers of ω in a nontrivial manner, but in fact they cancel out. To see this, note that if we make the replacement $\omega \to \omega^2$ in (5.11), then $r_0 \to r_0,\ r_1 \to r_2,\ r_2 \to r_1$. Thus the effect of this replacement is merely to permute the roots r_1, r_2 of the expression $P_3(x) = (x - r_0)(x - r_1)(x - r_2)$, hence to leave $P_3(x)$ invariant. This means that $B = C$, and

$P_3(x) = A + B(\omega + \omega^2)$. However, $1 + \omega + \omega^2 = 0$ so $P_3(x) = A - B$. Working out the details we find

$$P_3(x) = x^3 - 3z_0x^2 + 3(z_0^2 - z_1z_2)x + (3z_0z_1z_2 - z_0^3 - z_1^3 - z_2^3) = x^3 + ax^2 + bx + c.$$

Comparing coefficients of powers of x we obtain the identities

$$z_0 = -\frac{a}{3}, \quad z_1z_2 = \frac{a^2 - 3b}{9}, \quad z_1^3 + z_2^3 = \frac{-2a^3 + 9ab - 27c}{27},$$

or

$$z_1^3z_2^3 = \left(\frac{a^2 - 3b}{9}\right)^3, \quad z_1^3 + z_2^3 = \frac{-2a^3 + 9ab - 27c}{27}.$$

It is simple algebra to solve these two equations for the unknowns z_1^3, z_2^3:

$$z_1^3 = \frac{-2a^3 + 9ab - 27c}{54} + D, \quad z_2^3 = \frac{-2a^3 + 9ab - 27c}{54} - D,$$

where

$$D^2 = \left(\frac{-2a^3 + 9ab - 27c}{54}\right)^2 - \left(\frac{a^2 - 3b}{9}\right)^3.$$

Taking cube roots, we can obtain z_1, z_2 and plug the solutions for z_0, z_1, z_2 back into (5.11) to arrive at the Cardan formulas for r_0, r_1, r_2.

This method also works for finding the roots of second order polynomials (where it is trivial) and fourth order polynomials (where it is much more complicated). Of course it, and all such explicit methods, must fail for fifth and higher order polynomials.

5.3 Fast Fourier Transform (FFT)

We have shown that DFT for the column N-vector $\{f[n]\}$ is determined by the equation $F = \mathcal{F}_N f$, or in detail,

$$F[k] = \sum_{n=0}^{N-1} f[n]e_k[-n] = \sum_{n=0}^{N-1} f[n]\omega^{kn}, \qquad \omega = e^{-2\pi i/N}.$$

¿From this equation we see that each computation of the N-vector F requires N^2 multiplications of complex numbers. However, due to the special structure of the matrix $\mathcal{F}_N$ we can greatly reduce the number of multiplications and speed up the calculation. (We will ignore the number of additions, since they can be done much faster and with much less computer memory.) The procedure for doing this is the Fast Fourier Transform (FFT). It reduces the number of multiplications to about $N\log_2 N$.

The algorithm requires $N = 2^M$ for some integer M, so $F = \mathcal{F}_{2^M} f$. We split f into its even and odd components:

$$F[k] = \sum_{n=0}^{N/2-1} f[2n]\omega^{2kn} + \omega^k \sum_{n=0}^{N/2-1} f[2n+1]\omega^{2kn}.$$

Note that each of the sums has period $N/2 = 2^{M-1}$ in k, so

$$F[k+N/2] = \sum_{n=0}^{N/2-1} f[2n]\omega^{2kn} - \omega^k \sum_{n=0}^{N/2-1} f[2n+1]\omega^{2kn}.$$

Thus by computing the sums for $k = 0, 1\cdots, N/2-1$, hence computing $F[k]$ we get the $F[k+N/2]$ virtually for free. Note that the first sum is the DFT of the downsampled f and the second sum is the DFT of the data vector obtained from f by first left shifting and then downsampling.

Let's rewrite this result in matrix notation. We split the $N = 2^M$ component vector f into its even and odd parts, the 2^{M-1}-vectors

$$f_e = (f[0], f[2], \cdots, f[N-2]), \qquad f_o = (f[1], f[3], \cdots, f[N-1]),$$

and divide the N-vector F into halves, the $N/2$-vectors

$$F_- = (F[0], F[1], \cdots, F[N/2-1]), \qquad F_+ = (F[0+N/2], F[1+N/2], \cdots, F[N-1]).$$

We also introduce the $N/2 \times N/2$ diagonal matrix $D_{N/2}$ with matrix elements $(D_{N/2})_{jk} = \omega^k \delta_{jk}$, and the $N/2 \times N/2$ zero matrix $(O_{N/2})_{jk} = 0$ and identity matrix $(I_{N/2})_{jk} = \delta_{jk}$. The above two equations become

$$F_- = \mathcal{F}_{N/2} f_e + D_{N/2}\mathcal{F}_{N/2} f_o, \qquad F_+ = \mathcal{F}_{N/2} f_e - D_{N/2}\mathcal{F}_{N/2} f_o,$$

or

$$\begin{pmatrix} F_- \\ F_+ \end{pmatrix} = \begin{pmatrix} I_{N/2} & D_{N/2} \\ I_{N/2} & -D_{N/2} \end{pmatrix} \begin{pmatrix} \mathcal{F}_{N/2} & O_{N/2} \\ O_{N/2} & \mathcal{F}_{N/2} \end{pmatrix} \begin{pmatrix} f_e \\ f_0 \end{pmatrix} \tag{5.12}$$

Note that this factorization of the transform matrix has reduced the number of multiplications for the DFT from 2^{2M} to $2^{2M-1} + 2^M$, i.e., cut them about in half, for large M. We can apply the same factorization technique to $\mathcal{F}_{N/2}$, $\mathcal{F}_{N/4}$, and so on, iterating M times. Each iteration involves 2^M multiplications, so the total number of FFT multiplications is about $M2^M$. Thus a $N = 2^{10} = 1024$ point DFT which originally involved $N^2 = 2^{20} = 1,048,576$ complex multiplications, can be computed via the FFT with $10 \times 2^{10} = 10,240$ multiplications. In addition to the very impressive speed-up in the computation, there is an improvement in accuracy. Fewer multiplications leads to a smaller roundoff error.

5.4 Approximation to the Fourier Transform

One of the most important applications of the FFT is to the approximation of Fourier transforms. We will indicate how to set up the calculation of the FFT to calculate the Fourier transform of a continuous function $f(t)$ with support on the interval $\alpha \leq t \leq \beta$ of the real line. We first map this interval on the line to the unit circle $0 \leq \phi \leq 2\pi$, mod 2π via the affine transformation $t = a\phi + b$. Clearly $b = \alpha, a = \frac{\beta-\alpha}{2\pi}$. Since normally $f(\beta) \neq f(\alpha)$ when we transfer f as a function on the unit circle there will usually be a jump discontinuity at β. Thus we can expect the Gibbs phenomenon to occur there.

We want to approximate

$$\hat{f}(\lambda) = \int_{-\infty}^{\infty} f(t)e^{-it\lambda}dt = \int_{\alpha}^{\beta} f(t)e^{-it\lambda}dt$$

$$= \frac{\beta-\alpha}{2\pi}e^{-i\lambda\alpha}\int_0^{2\pi} g(\phi)e^{-i\frac{\beta-\alpha}{2\pi}\lambda\phi}d\phi$$

$$= \frac{\beta-\alpha}{4\pi^2}e^{-i\lambda\alpha}\hat{g}(\frac{\beta-\alpha}{2\pi}\lambda)$$

where $g(\phi) = f(\frac{\beta-\alpha}{2\pi}\phi + \alpha)$.

For an N-vector DFT we will choose our sample points at $\phi = \frac{2\pi n}{N}, \quad n = 0, 1, \cdots, N-1$. Thus

$$g = (g[0], g[1], \cdots, g[N-1]), \qquad g[n] = f(\frac{\beta-\alpha}{N}n + \alpha).$$

Now the Fourier coefficients

$$G[k] = \sum_{n=0}^{N-1} g[n]e_k[-n] = \sum_{n=0}^{N-1} g[n]\omega^{kn}, \qquad \omega = e^{-2\pi i/N}$$

are approximations of the coefficients $\hat{g}(k)$. Indeed $\hat{g}(k) \sim G[k]/N$. Thus

$$\hat{f}(\frac{2\pi k}{\beta - \alpha}) \sim \frac{\beta - \alpha}{4\pi^2 N} e^{\frac{-2\pi i k \alpha}{\beta - \alpha}} G[k].$$

Note that this approach is closely related to the ideas behind the Shannon sampling theorem, except here it is the signal $f(t)$ that is assumed to have compact support. Thus $f(t)$ can be expanded in a Fourier series on the interval $[\alpha, \beta]$ and the DFT allows us to approximate the Fourier series coefficients from a sampling of $f(t)$. (This approximation is more or less accurate, depending on the aliasing error.) Then we notice that the Fourier series coefficients are proportional to an evaluation of the Fourier transform $\hat{f}(\lambda)$ at the discrete points $\lambda = \frac{2\pi k}{\beta - \alpha}$ for $k = 0, \cdots, N - 1$.

Chapter 6

Linear Filters

In this chapter we will introduce and develop those parts of linear filter theory that are most closely related to the mathematics of wavelets, in particular perfect reconstruction filter banks. We will primarily, though not exclusively, be concerned with discrete filters. I will modify some of my notation for vector spaces, and Fourier transforms so as to be in accordance with the text by Strang and Nguyen.

6.1 Discrete Linear Filters

A *discrete-time signal* is a sequence of numbers (real or complex, but usually real). The signal takes the form

$$\mathbf{x} = (\cdots, x_{-1}, x_0, x_1, x_2, \cdots) \qquad \text{or} \quad \mathbf{x} = \begin{bmatrix} \vdots \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \mathbf{x}(2) \\ \vdots \end{bmatrix}.$$

Intuitively, we think of $\mathbf{x}(n)$ as the signal at time nT where T is the time interval between successive signals. $\mathbf{x}$ could be a digital sampling of a continuous analog signal or simply a discrete data stream. In general, these signals are of infinite length. (Later, we will consider signals of fixed finite length.) Usually, but not always, we will require that the signals belong to ℓ^2, i.e., that they have finite

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}(-1) \\ \mathbf{y}(0) \\ \mathbf{y}(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & \mathbf{h}(0) & \mathbf{h}(-1) & \mathbf{h}(-2) & \cdot & \cdot \\ \cdot & \cdot & \mathbf{h}(1) & \mathbf{h}(0) & \mathbf{h}(-1) & \cdot & \cdot \\ \cdot & \cdot & \mathbf{h}(2) & \mathbf{h}(1) & \mathbf{h}(0) & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

Figure 6.1: Matrix filter action

energy: $\sum_{n=-\infty}^{\infty} |x_n|^2 < \infty$. Recall that ℓ^2 is a Hilbert space with inner product

$$(\mathbf{x}, \mathbf{y}) = \sum_{n=-\infty}^{\infty} x_n \bar{y}_n.$$

The impulses $\mathbf{e}_j, \quad j = 0, \pm 1, \pm 2, \cdots$ defined by $\mathbf{e}_j(n) = \delta_{jn}$ form an ON basis for the signal space: $(\mathbf{e}_j, \mathbf{e}_k) = \delta_{jk}$ and $\mathbf{x} = \sum_{n=-\infty}^{\infty} x_n \mathbf{e}_n$. In particular the impulse at time 0 is called the *unit impulse* $\delta = \mathbf{e}_0$. The right *shift* or *delay* operator $\mathbf{S} : \ell^2 \to \ell^2$ is defined by $\mathbf{S}\mathbf{x}(n) = \mathbf{x}(n-1)$. Note that the action of this bounded operator is to delay the signal by one unit. Similarly the inverse operator $\mathbf{S}^{-1}\mathbf{x}(n) = \mathbf{x}(n+1)$ advances the signal by one time unit.

A *digital filter* $\mathbf{H}$ is a bounded linear operator $\mathbf{H} : \ell^2 \to \ell^2$ that is time invariant. The filter processes each input $\mathbf{x}$ and gives an output $\mathbf{H}\mathbf{x} = \mathbf{y}$. Since $\mathbf{H}$ is linear, its action is completely determined by the outputs $\mathbf{H}\mathbf{e}_j$. Time invariance means that $\mathbf{H}\mathbf{x}^k = \mathbf{y}^k, \ k = 0, \pm 1, \cdots$ whenever $\mathbf{H}\mathbf{x} = \mathbf{y}$. (Here, $\mathbf{x}^k(n) = \mathbf{x}(n-k)$.) Thus, the effect of delaying the input by k units of time is just to delay the output by k units. (Another way to put this is $\mathbf{HS} = \mathbf{SH}$, the filter commutes with shifts.) We can associate an infinite matrix with $\mathbf{H}$.

$$\mathbf{H} = [H_{mn}], \qquad \text{where } H_{mn} = (\mathbf{H}\mathbf{e}_n, \mathbf{e}_m).$$

Thus, $\mathbf{H}\mathbf{e}_n = \sum_{m=-\infty}^{\infty} H_{mn}\mathbf{e}_m$ and $\mathbf{y}(m) = \sum_{n=-\infty}^{\infty} H_{mn}\mathbf{x}(n)$. In terms of the matrix elements, time invariance means $H_{mn} = (\mathbf{H}\mathbf{e}_n, \mathbf{e}_m) = (\mathbf{H}\mathbf{e}_{n+k}, \mathbf{e}_{m+k}) = H_{m+k,n+k}$ for all k . Hence The matrix elements H_{mn} depend only on the difference $m-n$: $H_{mn} = \mathbf{h}(m-n)$ and $\mathbf{H}$ is completely determined by its coefficients $\mathbf{h}(j)$. The filter action looks like Figure 6.1. Note that the matrix has diagonal bands. $\mathbf{h}(0)$ appears down the main diagonal, $\mathbf{h}(-1)$ on the first superdiagonal, $\mathbf{h}(-2)$ on the next superdiagonal, etc. Similarly $\mathbf{h}(1)$ on the first subdiagonal, etc.

A matrix H_{mn} whose matrix elements depend only on $n - m$ is called a *Toeplitz matrix*.

Another way that we can express the action of the filter is in terms of the shift operator:

$$\mathbf{H} = \sum_{n=-\infty}^{\infty} \mathbf{h}(n)\mathbf{S}^n. \tag{6.1}$$

Thus $\mathbf{s}^n\mathbf{e}_j = \mathbf{e}_{j+n}$ and

$$\mathbf{H}(\sum \mathbf{x}(j)\mathbf{e}_j) = \sum_{n,j} \mathbf{x}(j)\mathbf{h}(n)\mathbf{e}_{j+n} = \sum_{m,j} \mathbf{x}(j)\mathbf{h}(m-j)\mathbf{e}_m = \sum \mathbf{y}(m)\mathbf{e}_m.$$

We have $\mathbf{y}(m) = \sum_j \mathbf{x}(j)\mathbf{h}(m-j)$. If only a finite number of the coefficients $\mathbf{h}(n)$ are nonzero we say that we have a *Finite Impulse Response* (FIR) filter. Otherwise we have an *Infinite Impulse Response* (IIR) filter. We can uniquely define the action of an FIR filter on *any* sequence $\mathbf{x}$, not just an element of ℓ^2 because there are only a finite number of nonzero terms in (6.1), so no convergence difficulties. For an IIR filter we have to be more careful. Note that the response to the unit impulse is $\mathbf{H}\delta = \sum_n \mathbf{h}(n)\mathbf{e}_n$.

Finally, we say that a digital filter is *causal* if it doesn't respond to a signal until the signal is received, i.e., $\mathbf{h}(n) = 0$ for $n < 0$. To sum up, a causal FIR digital filter is completely determined by the impulse response vector

$$\mathbf{h} = (\mathbf{h}(0), \mathbf{h}(1), \cdots, \mathbf{h}(N))$$

where N is the largest nonnegative integer such that $\mathbf{h}(N) \neq 0$. We say that the filter has $N + 1$ "taps".

There are other ways to represent the filter action that will prove very useful. The next of these is in terms of the convolution of vectors. Recall that a signal $\mathbf{x}$ belongs to the Banach space ℓ^1 provided $\sum_{n=-\infty}^{\infty} |\mathbf{x}(n)| < \infty$.

Definition 27 *Let $\mathbf{x}, \mathbf{y}$ be in ℓ^1. The convolution $\mathbf{x} * \mathbf{y}$ is given by the expression*

$$\mathbf{x} * \mathbf{y}(n) = \sum_{m=-\infty}^{\infty} \mathbf{x}(n-m)\mathbf{y}(m), \qquad n = 0, \pm 1, \pm 2, \cdots.$$

Lemma 33

1. $\mathbf{x} * \mathbf{y} \in \ell^1$

2. $\mathbf{x} * \mathbf{y} = \mathbf{y} * \mathbf{x}$

SKETCH OF PROOF:

$$\sum_n |\mathbf{x} * \mathbf{y}(n)| \leq \sum_n \sum_m |\mathbf{x}(n-m)| \cdot |\mathbf{y}(m)| = \sum_m \sum_n |\mathbf{x}(n)| \cdot |\mathbf{y}(m)|$$

$$= (\sum_n |\mathbf{x}(n)|)(\sum_m |\mathbf{y}(m)|) < \infty.$$

The interchange of order of summation is justified because the series are absolutely convergent. Q.E.D.

REMARK: It is easy to show that if $\mathbf{x} \in \ell^1$ then $\mathbf{x} \in \ell^2$.

Now we note that the action of $\mathbf{H}$ can be given by the convolution

$$\mathbf{H}\mathbf{x} = \mathbf{h} * \mathbf{x}.$$

For an FIR filter, this expression makes sense even if $\mathbf{x}$ isn't in ℓ^1, since the sum is finite.

6.2 Continuous filters

The emphasis in this course will be on discrete filters, but we will examine a few basic definitions and concepts in the theory of continuous filters.

A *continuous-time signal* is a function $f(t)$ (real or complex-valued, but usually real), defined for all t on the real line . Intuitively, we think of $f(t)$ as the signal at time t, say a continuous analog signal. In general, these signals are of infinite length. Usually, but not always, we will require that the signals belong to $L^2[-\infty, \infty]$, i.e., that they have finite energy: $\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty$. Sometimes we will require that $f \in L^1[\infty, \infty]$.

The *time-shift* operator $\mathbf{S}_a : L^2 \to L^2$ is defined by $\mathbf{S}_a f(t) = f(t-a)$. The action of this bounded operator is to delay the signal by the time interval a. Similarly the inverse operator $\mathbf{S}_a^{-1} f(t) = f(t+a)$ advances the signal by a time units.

A *continuous filter* $\mathbf{H}$ is a bounded linear operator $\mathbf{H} : L^2 \to L^2$ that is time invariant. The filter processes each input f and gives an output $\mathbf{H}f = g$. Time invariance means that $\mathbf{H}(\mathbf{S}_a f)(t) = \mathbf{S}_a g(t)$, whenever $\mathbf{H}f(t) = g(t)$. Thus, the effect of delaying the input by a units of time is just to delay the output by a units. (Another way to put this is $\mathbf{H}\mathbf{S}_a = \mathbf{S}_a\mathbf{H}$, the filter commutes with shifts.)

Suppose that $\mathbf{H}$ takes the form

$$\mathbf{H}f(t) = \int_{-\infty}^{\infty} K(s,t) f(s) ds$$

where $K(s,t)$ is a continuous function in the (s,t) plane and with bounded support. Then the time invariance requirement

$$g(t-a) = \int_{-\infty}^{\infty} K(s,t) f(s-a) ds$$

whenever

$$g(t) = \int_{-\infty}^{\infty} K(s,t) f(s) ds,$$

for all $f \in L^2[-\infty,\infty]$ and all $a \in R$ implies $K(s+a,t+a) = K(s,t)$ for all real s,t; hence there is a continuous function h on the real line such that $K(s,t) = h(t-s)$. It follows that $\mathbf{H}$ is a convolution operator, i.e.,

$$\mathbf{H}f(t) = h * f(t) = f * h(t) = \int_{-\infty}^{\infty} h(s) f(t-s) ds$$

(Note: This characterization of a continuous filter as a convolution can be proved under rather general circumstances. However, h may not be a continuous function. Indeed for the identity operator, $h(s) = \delta(s)$, the Dirac delta function.)

Finally, we say that a continuous filter is *causal* if it doesn't respond to a signal until the signal is received, i.e., $\mathbf{H}f(t) = 0$ for $t < 0$ if $f(t) = 0$ for $t < 0$. This implies $h(s) = 0$ for $s < 0$. Thus a causal filter is completely determined by the impulse response function $h(s), s \geq 0$ and we have

$$\mathbf{H}f(t) = h * f(t) = f * h(t) = \int_{0}^{\infty} h(s) f(t-s) ds = \int_{-\infty}^{t} h(t-s) f(s) ds$$

If $f, h \in L^1[-\infty,\infty]$ then $g = \mathbf{H}f \in L^1[-\infty,\infty]$ and, by the convolution theorem

$$\hat{g}(\lambda) = \hat{h}(\lambda)\hat{f}(\lambda).$$

6.3 Discrete filters in the frequency domain: Fourier series and the Z-transform

Let $\mathbf{x} \in \ell^2$ be a discrete-time signal.

$$\mathbf{x} = (\cdots, x_{-1}, x_0, x_1, x_2, \cdots) \qquad \text{or} \quad \mathbf{x} = \begin{bmatrix} \vdots \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \mathbf{x}(2) \\ \vdots \end{bmatrix}.$$

Definition 28 *The discrete-time Fourier transform of* $\mathbf{x}$ *is*

$$X(\omega) = \sum_{n=-\infty}^{\infty} \mathbf{x}(n)e^{-in\omega}.$$

Note the change in point of view. The input is the set of coefficients $\mathbf{x}(n)$ and the output is the 2π-periodic function $X(\omega) \in L^2[-\pi, \pi]$. We consider $X(\omega)$ as the frequency-domain signal. We can recover the time domain signal from the frequency domain signal by integrating:

$$\mathbf{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)e^{in\omega}d\omega, \qquad n = 0, \pm 1, \cdots.$$

For discrete-time signals $\mathbf{x}$, $\mathbf{y}$ the Parseval identity is

$$(\mathbf{x}, \mathbf{y}) = \sum_{n=-\infty}^{\infty} \mathbf{x}(n)\overline{\mathbf{y}}(n) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)\overline{Y}(\omega)d\omega.$$

If $\mathbf{x}$ belongs to ℓ^1 then the Fourier transform X is a bounded continuous function on $[-\pi, \pi]$.

In addition to the *mathematics* notation $X(\omega)$ for the frequency-domain signal, we shall sometimes use the *signal processing* notation

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} \mathbf{x}(n)e^{-jn\omega}, \qquad j = \sqrt{-1}, \tag{6.2}$$

and the *z-transform* notation

$$X(z) = \sum_{n=-\infty}^{\infty} \mathbf{x}(n)z^{-n}, \tag{6.3}$$

Note that the z-transform is a function of the complex variable z. It reduces to the signal processing form for $z = e^{j\omega}$. The Fourier transform of the impulse response function $\mathbf{h}$ of an FIR filter is a polynomial in z^{-1}.

We need to discuss what high frequency and low frequency mean in the context of discrete-time signals. We try a thought experiment. We would want a constant signal $\mathbf{x}(n) \equiv 1$ to have zero frequency and it corresponds to $X(\omega) = \delta(\omega - 0)$ where $\delta(\omega)$ is the Dirac Delta Function, so $\omega = 0$ corresponds to low frequency. The highest possible degree of oscillation for a discrete-time signal would be $\mathbf{x}(n) = (-1)^n$, i.e., the signal changes sign in each successive time interval. This corresponds to the time-domain signal $X(\omega) = \delta(\omega - \pi)$. Thus $\pm\pi$, and not 2π, correspond to high frequency.

We can clarify this question further by considering two examples that do belong to the space ℓ^2. Consider first the discrete signal $\mathbf{x}^{(1)}$ where

$$\mathbf{x}^{(1)}[n] = \begin{cases} 1 & \text{if } -N \leq n \leq N \\ 0 & \text{otherwise.} \end{cases}$$

If N is a large integer then this signal is a nonzero constant for a long period, and there are only two discontinuities. Thus we would expect this signal to be (mostly) low frequency. The Fourier transform is, making use of the derivation (3.10) of the kernel function $D_k(x)$,

$$X^{(1)}(\omega) = \sum_{n=-\infty}^{\infty} \mathbf{x}^{(1)}[n]e^{-i\omega n} = \sum_{n=-N}^{N} e^{-i\omega n}$$

$$= \frac{\sin([N+1/2]\omega)}{\sin(\omega/2)} = 2D_N(\omega).$$

As we have seen, this function has a sharp maximum of $2N+1$ at $\omega = 0$ and falls off rapidly for $|\omega| > 0$.

Our second example is $\mathbf{x}^{(2)}$ where

$$\mathbf{x}^{(2)}[n] = \begin{cases} (-1)^n & \text{if } -N \leq n \leq N \\ 0 & \text{otherwise.} \end{cases}$$

If N is a large integer this signal oscillates as rapidly as possible for an extended period. Thus we would expect this signal to exhibit high frequency behavior. The Fourier transform is, with a small modification of our last calculation,

$$X^{(2)}(\omega) = \sum_{n=-\infty}^{\infty} \mathbf{x}^{(1)}[n]e^{-i\omega n} = \sum_{n=-N}^{N} (-1)^n e^{-i\omega n}$$

$$= (-1)^N \frac{\cos([N+1/2]\omega)}{\cos(\omega/2)} = 2D_N(\omega+\pi).$$

This function has a sharp maximum of $2N+1$ at $\omega = \pi$. It is clear that $\omega = 0, \pi, \bmod 2\pi$ correspond to low and high frequency, respectively.

In analogy with the properties of convolution for the Fourier transform on $[-\infty, \infty]$ we have the

Lemma 34 *Let* $\mathbf{x}, \mathbf{y}$ *be in* ℓ^1 *with frequency domain transforms* $X(\omega), Y(\omega)$, *respectively. The frequency-domain transform of the convolution* $\mathbf{x}*\mathbf{y}$ *is* $X(\omega)Y(\omega)$.

PROOF:

$$\sum_n \mathbf{x} * \mathbf{y}(n)e^{-in\omega} = \sum_n \sum_m \mathbf{x}(n-m)\cdot \mathbf{y}(m)e^{-i(n-m)\omega}e^{-im\omega}$$

$$= \sum_m (\sum_n \mathbf{x}(n-m)e^{-i(n-m)\omega})\mathbf{y}(m)e^{-im\omega}$$

$$= \sum_m (\sum_n \mathbf{x}(n)e^{-in\omega})\mathbf{y}(m)e^{-im\omega} = X(\omega)Y(\omega).$$

The interchange of order of summation is justified because the series converge absolutely. Q.E.D.

NOTE: If $\mathbf{h}$ has only a *finite* number of nonzero terms and $\mathbf{x} \in \ell^2$ (but not ℓ^1) then the interchange in order of summation is still justified in the above computation, and the transform of $\mathbf{h} * \mathbf{x}$ is $H(\omega)X(\omega)$.

Let $\mathbf{H}$ be a digital filter: $\mathbf{Hx} = \mathbf{h} * \mathbf{x}$. If $\mathbf{h} \in \ell^1$ then we have that the action of $\mathbf{H}$ in the frequency domain is given by

$$X(\omega) \to H(\omega)X(\omega)$$

where $H(\omega)$ is the frequency transform of $\mathbf{h}$. If $\mathbf{H}$ is a FIR filter then

$$X(z) \to H(z)X(z)$$

where $H(z)$ is a *polynomial* in z^{-1}.

One of the principal functions of a filter is to select a band of frequencies to pass, and to reject other frequencies. In the *pass band* $|H(\omega)|$ is maximal (or very close to its maximum value). We shall frequently normalize the filter so that this maximum value is 1. In the *stop band* $|H(\omega)|$ is 0, or very close to 0. Mathematically ideal filters can divide the spectrum into pass band and stop band; for realizable, non-ideal, filters there is a transition band where $|H(\omega)|$ changes from near 1 to near 0. A *low pass* filter is a filter whose passband is a band of frequencies around $\omega = 0$. (Indeed in this course we shall additionally require $H(0) = 1$ and $H(\pi) = 0$ for a low pass filter. Thus, if $\mathbf{H}$ is an FIR low pass filter we have $H(0) = \sum_{n=0}^{N} \mathbf{h}(n) = 1$.) A *high pass* filter is a filter whose pass band is a band of frequencies around $\omega = \pi$, (and in this course we shall additionally require $|H(\pi)| = 1$ and $H(0) = 0$ for a high pass filter.)

EXAMPLES:

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}(-1) \\ \mathbf{y}(0) \\ \mathbf{y}(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdot & \cdot \\ \cdot & \cdot & \frac{1}{2} & \frac{1}{2} & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & \frac{1}{2} & \frac{1}{2} & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

Figure 6.2: Moving average filter action

1. A simple low pass filter (moving average). *This is a very important example, associated with the Haar wavelets.* $\mathbf{Hx} = \mathbf{y}$ where $\mathbf{y}(n) = \frac{1}{2}\mathbf{x}(n) + \frac{1}{2}\mathbf{x}(n-1)$. $N = 1$ and the filter coefficients are $\mathbf{h} = (\mathbf{h}(0), \mathbf{h}(1)) = (\frac{1}{2}, \frac{1}{2})$. An alternate representation is $H = \frac{1}{2}\mathbf{I} + \frac{1}{2}\mathbf{S}$. The frequency response is $H(\omega) = \frac{1}{2} + \frac{1}{2}e^{-i\omega} = |H(\omega)|e^{i\phi(\omega)}$ where
$$|H(\omega)| = \cos\frac{\omega}{2}, \qquad \phi(\omega) = -\frac{\omega}{2}.$$
Note that $|H(\omega)|$ is 1 for $\omega = 0$ and 0 for $\omega = \pi$. This is a low pass filter. The z-transform is $H(z) = \frac{1}{2} + \frac{1}{2}z^{-1}$. The matrix form of the action in the time domain is given in Figure 6.2.

2. A simple high pass filter (moving difference). *This is also a very important example, associated with the Haar wavelets.* $\mathbf{Hx} = \mathbf{y}$ where $\mathbf{y}(n) = \frac{1}{2}\mathbf{x}(n) - \frac{1}{2}\mathbf{x}(n-1)$. $N = 1$ and the filter coefficients are $\mathbf{h} = (\mathbf{h}(0), \mathbf{h}(1)) = (\frac{1}{2}, -\frac{1}{2})$. An alternate representation is $H = \frac{1}{2}\mathbf{I} - \frac{1}{2}\mathbf{S}$. The frequency response is $H(\omega) = \frac{1}{2} - \frac{1}{2}e^{-i\omega} = |H(\omega)|e^{i\phi(\omega)}$ where
$$|H(\omega)| = \sin\frac{\omega}{2}, \qquad \phi(\omega) = \frac{\pi}{2} - \frac{\omega}{2}.$$
Note that $|H(\omega)|$ is 1 for $\omega = \pi$ and 0 for $\omega = 0$. This is a high pass filter. The z-transform is $H(z) = \frac{1}{2} - \frac{1}{2}z^{-1}$. The matrix form of the action in the time domain is given in Figure 6.3.

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}(-1) \\ \mathbf{y}(0) \\ \mathbf{y}(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & -\frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdot & \cdot \\ \cdot & \cdot & -\frac{1}{2} & \frac{1}{2} & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & -\frac{1}{2} & \frac{1}{2} & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

Figure 6.3: Moving difference filter action

6.4 Other operations on discrete signals in the time and frequency domains

We have already examined the action of a digital filter $\mathbf{H}$ in both the time and frequency domains. We will now do the same for some other useful operators. Let $\{\mathbf{x}(n) : n = 0, \pm 1, \cdots\}$ be a discrete-time signal in ℓ^2 with z-transform $X(z) = \sum_{n=-\infty}^{\infty} \mathbf{x}(n) z^{-n}$ in the time domain.

- Delay. $\mathbf{S}\mathbf{x}(n) = \mathbf{x}(n-1)$. In the frequency domain $\mathbf{S} : X(z) \rightarrow z^{-1}X(z)$, because

$$\sum_{n=-\infty}^{\infty} \mathbf{S}\mathbf{x}(n) z^{-n} = \sum_{n=-\infty}^{\infty} \mathbf{x}(n-1) z^{-n} = \sum_{m=-\infty}^{\infty} \mathbf{x}(m) z^{-m-1} = z^{-1}X(z).$$

- Advance. $\mathbf{S}^{-1}\mathbf{x}(n) = \mathbf{x}(n+1)$. In the frequency domain $\mathbf{S}^{-1} : X(z) \rightarrow zX(z)$.

- Downsampling. $(\downarrow 2)\mathbf{x}(n) = \mathbf{x}(2n)$, i.e.,

$$(\downarrow 2)\mathbf{x} = (\cdots, \mathbf{x}(-2), \mathbf{x}(0), \mathbf{x}(2), \cdots).$$

  In terms of matrix notation, the action of downsampling in the time domain is given by Figure 6.4. In terms of the z-transform, the action is

$$\sum_n (\downarrow 2)\mathbf{x}(n) z^{-n} = \sum_n \mathbf{x}(2n) z^{-n} = \frac{1}{2}\sum_n \mathbf{x}(n)(z^{\frac{1}{2}})^{-n} + \frac{1}{2}\sum_n \mathbf{x}(n)(-z^{\frac{1}{2}})^{-n}$$

$$= \frac{1}{2}[X(z^{\frac{1}{2}}) + X(-z^{\frac{1}{2}})].$$

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{x}_1(-2) \\ \mathbf{x}_1(0) \\ \mathbf{x}_1(2) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & 1 & 0 & 0 & 0 & 0 & \cdot \\ \cdot & 0 & 0 & 1 & 0 & 0 & \cdot \\ \cdot & 0 & 0 & 0 & 0 & 1 & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \mathbf{x}(2) \\ \cdot \end{bmatrix}. \qquad (6.4)$$

Figure 6.4: Downsampling matrix action

$$\begin{bmatrix} \cdot \\ \mathbf{x}(-1) \\ 0 \\ \mathbf{x}(0) \\ 0 \\ \mathbf{x}(1) \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & \cdot & \cdot \\ \cdot & 1 & 0 & 0 & \cdot \\ \cdot & 0 & 0 & 0 & \cdot \\ \cdot & 0 & 1 & 0 & \cdot \\ \cdot & 0 & 0 & 0 & \cdot \\ \cdot & 0 & 0 & 1 & \cdot \\ \cdot & \cdot & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \end{bmatrix}. \qquad (6.5)$$

Figure 6.5: Upsampling matrix action

- Upsampling. $(\uparrow 2)\mathbf{x}(n) = \begin{cases} \mathbf{x}(\frac{n}{2}) & n \text{ even} \\ 0 & n \text{ odd} \end{cases}$, i.e.,

  $$(\uparrow 2)\mathbf{x} = (\cdots, \mathbf{x}(-1), 0, \mathbf{x}(0), 0, \mathbf{x}(1), 0, \cdots).$$

  In terms of matrix notation, the action of upsampling in the time domain is given by Figure 6.5. In terms of the z-transform, the action is

  $$\sum_n (\uparrow 2)\mathbf{x}(n) z^{-n} = \sum_m \mathbf{x}(m) z^{-2m} = X(z^2).$$

- Upsampling followed by downsampling. $(\downarrow 2)(\uparrow 2)\mathbf{x}(n) = \mathbf{x}(n)$, the identity operator. Note that the matrices (6.4), (6.5) give $(\downarrow 2)(\uparrow 2) = I$, the infinite identity matrix. This shows that the upsampling matrix is the right inverse of the downsampling matrix. Furthermore the upsampling matrix is just the transpose of the downsampling matrix: $(\downarrow 2)^{\text{tr}} = (\uparrow 2)$.

- Downsampling followed by upsampling. $(\uparrow 2)(\downarrow 2)\mathbf{x}(n) = \begin{cases} \mathbf{x}(n) & n \text{ even} \\ 0 & n \text{ odd} \end{cases}$, i.e.,

  $$(\uparrow 2)(\downarrow 2)\mathbf{x} = (\cdots, \mathbf{x}(-2), 0, \mathbf{x}(0), 0, \mathbf{x}(2), 0, \cdots).$$

Note that the matrices (6.5), (6.4), give $(\uparrow 2)(\downarrow 2) \neq I$. This shows that the upsampling matrix is *not* the left inverse of the downsampling matrix. The action in the frequency domain is

$$(\uparrow 2)(\downarrow 2) : X(z) \rightarrow \frac{1}{2}[X(z) + X(-z)].$$

- Flip about $N/2$. The action in the time domain is $\mathbf{F}_{N/2}\mathbf{x}(n) = \mathbf{x}(N-n)$, i.e., reflect $\mathbf{x}$ about $N/2$. If N is even then the point $\mathbf{x}(N/2)$ is fixed. In the frequency domain we have

$$\mathbf{F}_{N/2} : X(z) \rightarrow z^{-N}X(z^{-1}).$$

- Alternate signs. $\mathbf{A}\mathbf{x}(n) = (-1)^n\mathbf{x}(n)$ or

$$\mathbf{A}\mathbf{x} = (\cdots, -\mathbf{x}(-1), \mathbf{x}(0), -\mathbf{x}(1), \mathbf{x}(2), \cdots).$$

Here

$$\mathbf{A} : X(z) \rightarrow X(-z).$$

- Alternating flip about $N/2$. The action in the time domain is $\mathbf{F}_{N/2}\mathbf{x}(n) = (-1)^n\mathbf{x}(N-n)$. In the frequency domain

$$X(z) \rightarrow (-z)^{-N}X(-z^{-1}).$$

- Conjugate alternating flip about $N/2$. The action in the time domain is $\mathbf{F}_{N/2}\mathbf{x}(n) = (-1)^n\overline{\mathbf{x}}(N-n)$. In the frequency domain

$$X(z) \rightarrow (-z)^{-N}\overline{X(-\overline{z}^{-1})}.$$

6.5 Filter banks, orthogonal filter banks and perfect reconstruction of signals

I want to analyze signals $\mathbf{x}(n)$ with digital filters. For efficiency, it is OK to throw away some of the data generated by this analysis. However, I want to make sure that I don't (unintentionally) lose information about the original signal as I proceed with the analysis. Thus I want this analysis process to be invertible: I want to be able to recreate (synthesize) the signal from the analysis output. Further I

$$\begin{bmatrix} \cdot \\ \mathbf{y}_0(-2) \\ \mathbf{y}_0(-1) \\ \mathbf{y}_0(0) \\ \mathbf{y}_0(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{h}_0(0) & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{h}_0(1) & \mathbf{h}_0(0) & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{h}_0(2) & \mathbf{h}_0(1) & \mathbf{h}_0(0) & 0 & \cdot & \cdot \\ \cdot & \mathbf{h}_0(3) & \mathbf{h}_0(2) & \mathbf{h}_0(1) & \mathbf{h}_0(0) & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix},$$

Figure 6.6: $\mathbf{H}_0$ matrix action

want this synthesis process to be implemented by filters. Thus, if I link the input for the synthesis filters to the output of the analysis filters I should end up with the original signal except for a fixed delay of ℓ units caused by the processing in the filters: $\mathbf{x}(n-\ell)$. This is the basic idea of *Perfect Reconstruction* of signals.

If we try to carry out the analysis and synthesis with a single filter, it is essential that the filter be an invertible operator. A lowpass filter would certainly fail this requirement, for example, since it would screen out the high frequency part of the signal and lose all information about the high frequency components of $\mathbf{x}(n)$. For the time being, we will consider only FIR filters and the invertibility problem is even worse for this class of filters. Recall that the z-transform $H(z)$ of an FIR filter is a polynomial in z^{-1}. Now suppose that $\mathbf{H}$ is an invertible filter with inverse $\mathbf{H}^{-1}$. Since $\mathbf{H}\mathbf{H}^{-1} = \mathbf{I}$ where $\mathbf{I}$ is the identity filter, the convolution theorem gives us that

$$H(z)H^{-1}(z) = 1,$$

i.e., the z-transform of $\mathbf{H}^{-1}$ is the reciprocal of the z-transform of $\mathbf{H}$. Except for trivial cases the z-transform of $\mathbf{H}^{-1}$ cannot be a polynomial in z^{-1}. Hence if the (nontrivial) FIR filter has an inverse, it is *not* an FIR filter. Thus for perfect reconstruction with FIR filters, we will certainly need more than one filter.

Let's try a *filter bank* with two FIR filters, $\mathbf{H}_0$ and $\mathbf{H}_1$. The input is $\mathbf{x} = \{\mathbf{x}(n)\}$. The output of the filters is $\mathbf{y}_j = \mathbf{H}_j\mathbf{x}$, $j = 1, 2$.

The $\mathbf{H}_0$ filter action looks like Figure 6.6. and the $\mathbf{H}_1$ filter action looks like Figure 6.7. Note that each row of the infinite matrix H_0 contains all zeros, except for the terms $(\mathbf{h}_0(N), \mathbf{h}_0(N-1), \cdots, \mathbf{h}_0(0))$ which are shifted one column to the right for each successive row. Similarly, each row of the infinite matrix H_1 contains all zeros, except for the terms $(\mathbf{h}_1(N), \mathbf{h}_1(N-1), \cdots, \mathbf{h}_1(0))$ which are shifted one column to the right for each successive row. (We choose N to be the largest of N_0, N_1, where $\mathbf{H}_0$ has $N_0 + 1$ taps and $\mathbf{H}_1$ has $N_1 + 1$ taps.) Thus each

$$\begin{bmatrix} \cdot \\ \mathbf{y}_1(-2) \\ \mathbf{y}_1(-1) \\ \mathbf{y}_1(0) \\ \mathbf{y}_1(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{h}_1(0) & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{h}_1(1) & \mathbf{h}_1(0) & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{h}_1(2) & \mathbf{h}_1(1) & \mathbf{h}_1(0) & 0 & \cdot & \cdot \\ \cdot & \mathbf{h}_1(3) & \mathbf{h}_1(2) & \mathbf{h}_1(1) & \mathbf{h}_1(0) & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

Figure 6.7: $\mathbf{H}_1$ matrix action

$$\begin{bmatrix} \cdot \\ \mathbf{y}_0(-2) \\ \mathbf{y}_0(-1) \\ \mathbf{y}_0(0) \\ \mathbf{y}_0(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{c}(0) & 0 & 0 & 0\cdot & \cdot & \\ \cdot & \mathbf{c}(1) & \mathbf{c}(0) & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & 0 & \cdot & \cdot \\ \cdot & \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix},$$

Figure 6.8: $\mathbf{C}$ matrix action

row vector has the same norm $||\mathbf{h}_j||$.

It will turn out to be very convenient to have a filter all of whose row vectors have norm 1. Thus we will replace filter $\mathbf{H}_0$ by the normalized filter

$$\mathbf{C} = \frac{1}{||\mathbf{h}_0||}\mathbf{H}_0.$$

The impulse response vector for $\mathbf{C}$ is $\mathbf{c} = \frac{1}{||\mathbf{h}_0||}\mathbf{h}_0$, so that $||\mathbf{c}|| = 1$. Similarly, we will replace filter $\mathbf{H}_1$ by the normalized filter

$$\mathbf{D} = \frac{1}{||\mathbf{h}_1||}\mathbf{H}_1.$$

The impulse response vector for $\mathbf{D}$ is $\mathbf{d} = \frac{1}{||\mathbf{h}_1||}\mathbf{h}_1$, so that $||\mathbf{d}|| = 1$. The $\mathbf{C}$ filter action looks like Figure 6.8. and the $\mathbf{D}$ filter action looks like Figure 6.9.

Now these two filters are producing twice as much output as the original input, and we want eventually to compress the output (or certainly not add to the stream of data that is transmitted). Otherwise we would have to delay the data transmission by an ever growing amount, or we would have to replace the original

$$\begin{bmatrix} \cdot \\ \mathbf{y}_1(-2) \\ \mathbf{y}_1(-1) \\ \mathbf{y}_1(0) \\ \mathbf{y}_1(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{d}(0) & 0 & 0 & 0\cdot & \cdot & \\ \cdot & \mathbf{d}(1) & \mathbf{d}(0) & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{d}(2) & \mathbf{d}(1) & \mathbf{d}(0) & 0 & \cdot & \cdot \\ \cdot & \mathbf{d}(3) & \mathbf{d}(2) & \mathbf{d}(1) & \mathbf{d}(0) & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

Figure 6.9: $\mathbf{D}$ matrix action

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}_0(-2) \\ \mathbf{y}_0(0) \\ \mathbf{y}_0(2) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{c}(0) & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & 0 & \cdot & \cdot \\ \cdot & \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix},$$

Figure 6.10: $\mathbf{L}$ matrix action

one-channel transmission by a two-channel transmission. Thus we will downsample the output of filters $\mathbf{C}$ and $\mathbf{D}$. This will effectively replace or original filters $\mathbf{C}$ and $\mathbf{D}$ by new filters

$$\mathbf{L} = (\downarrow 2)\mathbf{C} = \frac{1}{||\mathbf{h}_0||}(\downarrow 2)\mathbf{H}_0, \quad \mathbf{B} = (\downarrow 2)\mathbf{D} = \frac{1}{||\mathbf{h}_1||}(\downarrow 2)\mathbf{H}_1.$$

The $\mathbf{L}$ filter action looks like Figure 6.10. and the $\mathbf{B}$ filter action looks like Figure 6.11. Note that each row vector is now shifted two spaces to the right of the row vector immediately above. Now we put the $\mathbf{L}$ and $\mathbf{B}$ matrices together to display the full time domain action $\mathbf{H}_t$ of the analysis part of this filter bank, see Figure 6.12. This is just the original full filter, with the odd-number rows removed. How can we ensure that this decimation of data from the original filters $\mathbf{C}$ and $\mathbf{D}$ still permits reconstruction of the original signal $\mathbf{x}$ from the truncated outputs $\{\mathbf{y}_0(2n)\}$,$\{\mathbf{y}_1(2n)\}$? The condition is, clearly, that the infinite matrix $\mathbf{H}_t$ should be invertable!

The invertability requirement is very strong, and won't be satisfied in general. For example, if $\mathbf{C}$ and $\mathbf{D}$ are both lowpass filters, then high frequency information

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}_1(-2) \\ \mathbf{y}_1(0) \\ \mathbf{y}_1(2) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{d}(0) & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{d}(2) & \mathbf{d}(1) & \mathbf{d}(0) & 0 & \cdot & \cdot \\ \cdot & \mathbf{d}(4) & \mathbf{d}(3) & \mathbf{d}(2) & \mathbf{d}(1) & \mathbf{d}(0) & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

Figure 6.11: $\mathbf{B}$ matrix action

$$\mathbf{H}_t = \begin{bmatrix} \mathbf{L} \\ \mathbf{B} \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{c}(0) & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & 0 & \cdot & \cdot \\ \cdot & \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{d}(0) & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{d}(2) & \mathbf{d}(1) & \mathbf{d}(0) & 0 & \cdot & \cdot \\ \cdot & \mathbf{d}(4) & \mathbf{d}(3) & \mathbf{d}(2) & \mathbf{d}(1) & \mathbf{d}(0) & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix}.$$

Figure 6.12: $\mathbf{H}_t$ matrix action

$$\overline{\mathbf{H}}_t^{\rm tr} = \left[\overline{\mathbf{L}}^{\rm tr} \quad \overline{\mathbf{B}}^{\rm tr}\right] = \left[\begin{array}{cccccccccc} \cdot & & & & \cdot & \cdot & & & & \cdot \\ \cdot & \overline{\mathbf{c}}(0) & \overline{\mathbf{c}}(2) & \overline{\mathbf{c}}(4) & \cdot & \cdot & \overline{\mathbf{d}}(0) & \overline{\mathbf{d}}(2) & \overline{\mathbf{d}}(4) & \cdot \\ \cdot & 0 & \overline{\mathbf{c}}(1) & \overline{\mathbf{c}}(3) & \cdot & \cdot & 0 & \overline{\mathbf{d}}(1) & \overline{\mathbf{d}}(3) & \cdot \\ \cdot & 0 & \overline{\mathbf{c}}(0) & \overline{\mathbf{c}}(2) & \cdot & \cdot & 0 & \overline{\mathbf{d}}(0) & \overline{\mathbf{d}}(2) & \cdot \\ \cdot & 0 & 0 & \overline{\mathbf{c}}(1) & \cdot & \cdot & 0 & 0 & \overline{\mathbf{d}}(1) & \cdot \\ \cdot & 0 & 0 & \overline{\mathbf{c}}(0) & \cdot & \cdot & 0 & 0 & \overline{\mathbf{d}}(0) & \cdot \\ \cdot & & & & \cdot & \cdot & & & & \cdot \end{array}\right].$$

Figure 6.13: $\overline{\mathbf{H}}_t^{\rm tr}$ matrix

from the original signal will be permanently lost. However, if $\mathbf{C}$ is a lowpass filter and $\mathbf{D}$ is a highpass filter, then there is hope that the high frequency information from $\mathbf{D}$ and the low frequency information from $\mathbf{C}$ will supplement one another, even after downsampling.

Initially we are going to make even a stronger requirement on $\mathbf{H}_t$ than invertability. We are going to require that $\mathbf{H}_t$ be a *unitary* matrix. (In that case the inverse of the matrix is just the transpose conjugate and solving for the original signal $\mathbf{x}$ from the truncated outputs $\{\mathbf{y}_0(2n)\}$,$\{\mathbf{y}_1(2n)\}$ is simple. Moreover, if the impulse response vectors $\mathbf{c}$, $\mathbf{d}$ are real, then the matrix will be *orthogonal*.)

The transpose conjugate looks like Figure 6.13.

UNITARITY CONDITION:

$$\overline{\mathbf{H}}_t^{\rm tr}\mathbf{H}_t = \mathbf{H}_t\overline{\mathbf{H}}_t^{\rm tr} = \mathbf{I}.$$

Written out in terms of the $\mathbf{L}$ and $\mathbf{B}$ matrices this is

$$\left[\overline{\mathbf{L}}^{\rm tr} \quad \overline{\mathbf{B}}^{\rm tr}\right] \left[\begin{array}{c} \mathbf{L} \\ \mathbf{B} \end{array}\right] = \overline{\mathbf{L}}^{\rm tr}\mathbf{L} + \overline{\mathbf{B}}^{\rm tr}\mathbf{B} = \mathbf{I} \tag{6.6}$$

and

$$\left[\begin{array}{c} \mathbf{L} \\ \mathbf{B} \end{array}\right] \left[\overline{\mathbf{L}}^{\rm tr} \quad \overline{\mathbf{B}}^{\rm tr}\right] = \left[\begin{array}{cc} \mathbf{L}\overline{\mathbf{L}}^{\rm tr} & \mathbf{L}\overline{\mathbf{B}}^{tr} \\ \mathbf{B}\overline{\mathbf{L}}^{\rm tr} & \mathbf{B}\overline{\mathbf{B}}^{\rm tr} \end{array}\right] = \left[\begin{array}{cc} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{array}\right]. \tag{6.7}$$

For the filter coefficients $\mathbf{c}(k)$ and $\mathbf{d}(k)$ conditions (6.7) become *orthogonality to double shifts* of the rows:

$$\mathbf{L}\mathbf{L}^{\rm tr} = \mathbf{I} : \qquad \sum_n \mathbf{c}(n)\overline{\mathbf{c}}(n-2k) = \delta_{k0} \tag{6.8}$$

$$\mathbf{LB}^{\mathrm{tr}} = \mathbf{0} : \qquad \sum_n \mathbf{c}(n)\overline{\mathbf{d}}(n-2k) = 0 \tag{6.9}$$

$$\mathbf{BB}^{\mathrm{tr}} = \mathbf{I} : \qquad \sum_n \mathbf{d}(n)\overline{\mathbf{d}}(n-2k) = \delta_{k0} \tag{6.10}$$

REMARKS ON THE UNITARITY CONDITION

- The condition says that the row vectors of $\mathbf{H}_t$ form an ON set, and that the column vectors of $\mathbf{H}_t$ also form an ON set. For a finite dimensional matrix, only one of these requirements is needed to imply orthogonality; the other property can be proved. For infinite matrices, however, both requirements are needed to imply orthogonality.

- By normalizing the rows of $\mathbf{H}_0$, $\mathbf{H}_1$ to length 1, hence replacing these filters by the normalized filters $\mathbf{C}$, $\mathbf{D}$ we have already gone part way to the verification of orthonormality.

- The double shift orthogonality conditions (6.8)-(6.10) force N to be odd. For if N were even, then setting $k = N/2$ in these equations (and also $k = -N/2$ in the middle one) leads to the conditions

  $$\mathbf{c}(N)\overline{\mathbf{c}}(0) = \mathbf{c}(N)\overline{\mathbf{d}}(0) = \mathbf{d}(N)\overline{\mathbf{c}}(0) = \mathbf{d}(N)\overline{\mathbf{d}}(0) = 0.$$

  This violates our definition of N.

- The orthogonality condition (6.8) says that the rows of $\mathbf{L}$ are orthogonal, and condition (6.10) says that the rows of $\mathbf{B}$ are orthogonal. Condition (6.9) says that the rows of $\mathbf{L}$ are orthogonal to the rows of $\mathbf{B}$.

- If we know that the rows of $\mathbf{L}$ are orthogonal, then we can alway construct a filter $\mathbf{B}$, hence the impulse response vector $\mathbf{d}$, such that conditions (6.9),(6.10) are satisfied. Suppose that $\mathbf{c}$ satisfies conditions (6.8). Then we define $\mathbf{d}$ by applying the conjugate alternating flip about $N/2$ to $\mathbf{c}$. (Recall that N *must be odd*. We are flipping the vector $\mathbf{c} = (\mathbf{c}(0), \mathbf{c}(1), \cdots, \mathbf{c}(N))$ about its midpoint, conjugating, and alternating the signs.)

  $$\mathbf{d}(n) = (-1)^n \overline{\mathbf{c}}(N-n), \qquad n = 0, 1, \cdots, N. \tag{6.11}$$

  Thus

  $$\mathbf{d} = (\mathbf{d}(0), \mathbf{d}(1), \cdots, \mathbf{d}(N)) = (\overline{\mathbf{c}}(N), -\overline{\mathbf{c}}(N-1), \overline{\mathbf{c}}(N-2), \cdots, -\overline{\mathbf{c}}(0))$$

$$\mathbf{H}_t = \left[\begin{array}{c} \mathbf{L} \\ \mathbf{B} \end{array}\right] = \left[\begin{array}{ccccccc}
\cdot & \cdot & & & & \cdot & \cdot \\
\cdot & \cdot & & & & \cdot & \cdot \\
\cdot & \mathbf{c}(0) & 0 & 0 & 0 & \cdot & \cdot \\
\cdot & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & 0 & \cdot & \cdot \\
\cdot & \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & \cdot \\
\cdot & \cdot & & & & \cdot & \cdot \\
\cdot & \cdot & & & & \cdot & \cdot \\
\cdot & \overline{\mathbf{c}}(N) & 0 & 0 & 0 & \cdot & \cdot \\
\cdot & \overline{\mathbf{c}}(N-2) & -\overline{\mathbf{c}}(N-1) & \overline{\mathbf{c}}(N) & 0 & \cdot & \cdot \\
\cdot & \overline{\mathbf{c}}(N-4) & -\overline{\mathbf{c}}(N-3) & \overline{\mathbf{c}}(N-2) & -\overline{\mathbf{c}}(N-1) & \overline{\mathbf{c}}(N) & \cdot \\
\cdot & \cdot & & & & \cdot & \cdot \\
\cdot & \cdot & & & & \cdot & \cdot
\end{array}\right].$$

Figure 6.14: $\mathbf{H}_t$ matrix

and $\mathbf{H}_t$ looks like Figure 6.14. You can check by taking simple examples that this works. However in detail:

$$S \equiv \sum_n \mathbf{c}(n)\overline{\mathbf{d}}(n-2k) = \sum_n \mathbf{c}(n)(-1)^n\mathbf{c}(N-n+2k).$$

Setting $m = N - n + 2k$ in the last sum we find

$$S = \sum_m \mathbf{c}(N-m+2k)\mathbf{c}(m)(-1)^{N-m} = -S,$$

since N is odd. Thus $S = 0$. Similarly,

$$T \equiv \sum_n \mathbf{d}(n)\overline{\mathbf{d}}(n-2k) = \sum_n (-1)^n\overline{\mathbf{c}}(N-n)(-1)^{n-2k}\mathbf{c}(N-n+2k)$$

$$= \sum_n \mathbf{c}(N-n+2k)\overline{\mathbf{c}}(N-n).$$

Now set $m = N - n + 2k$ in the last sum:

$$T = \sum_m \mathbf{c}(m)\overline{\mathbf{c}}(m-2k) = \delta_{k0}.$$

NOTE: This construction is no accident. Indeed, using the facts that $\mathbf{c}(0)\mathbf{c}(N) \neq 0$ and that the nonzero terms in a row of $\mathbf{B}$ overlap nonzero terms from a row of

$\mathbf{L}$ in exactly $0, 2, 4, \cdots, N+1$ places, you can derive that $\mathbf{d}$ must be related to $\mathbf{c}$ by ± a conjugate alternating flip, in order for the rows to be ON.

Now we have to consider the remaining condition (6.6), the orthonormality of the columns of $\mathbf{H}_t$. Note that the columns of $\mathbf{H}_t$ are of two types: even (containing only terms $\mathbf{c}(2n), \mathbf{d}(2n)$) and odd (containing only terms $\mathbf{c}(2n+1), \mathbf{d}(2n+1)$). Thus the requirement that the column vectors of $\mathbf{H}_t$ are ON reduces to 3 types of identities:

$$\begin{aligned}
\text{even} - \text{even} : \quad & \sum_\ell \mathbf{c}(2\ell)\overline{\mathbf{c}}(2k+2\ell) \\
+ \quad & \sum_\ell \mathbf{d}(2\ell)\overline{\mathbf{d}}(2k+2\ell) = \delta_{k0} \qquad (6.12) \\
\text{odd} - \text{odd} : \quad & \sum_\ell \mathbf{c}(2\ell+1)\overline{\mathbf{c}}(2k+2\ell+1) \\
+ \quad & \sum_\ell \mathbf{d}(2\ell+1)\overline{\mathbf{d}}(2k+2\ell+1) = \delta_{k0} \qquad (6.13) \\
\text{odd} - \text{even} : \quad & \sum_\ell \mathbf{c}(2\ell+1)\overline{\mathbf{c}}(2k+2\ell) \\
+ \quad & \sum_\ell \mathbf{d}(2\ell+1)\overline{\mathbf{d}}(2k+2\ell) = 0. \qquad (6.14)
\end{aligned}$$

Theorem 35 *If the filter $\mathbf{L}$ satisfies the double shift orthogonality condition (6.8) and the filter $\mathbf{B}$ is determined by the conjugate alternating flip*

$$\mathbf{d}(n) = (-1)^n \overline{\mathbf{c}}(N-n), \qquad n = 0, 1, \cdots, N,$$

then condition (6.6) holds and the columns of $\mathbf{H}_t$ are orthonormal.

PROOF:

1. even-even

$$\sum_\ell \mathbf{d}(2\ell)\overline{\mathbf{d}}(2k+2\ell) = \sum_\ell \overline{\mathbf{c}}(N-2\ell)\mathbf{c}(N-2k-2\ell)$$
$$= \sum_s \mathbf{c}(2s+1)\overline{\mathbf{c}}(2s+2k+1).$$

Thus

$$\sum_\ell \mathbf{c}(2\ell)\overline{\mathbf{c}}(2k+2\ell) + \sum_\ell \mathbf{d}(2\ell)\overline{\mathbf{d}}(2k+2\ell)$$
$$= \sum_n \mathbf{c}(n)\overline{\mathbf{c}}(n+2k) = \delta_{k0}$$

from (6.8).

2. odd-odd

$$\sum_\ell \mathbf{d}(2\ell+1)\overline{\mathbf{d}}(2k+2\ell+1) = \sum_\ell \overline{\mathbf{c}}(N-2\ell-1)\mathbf{c}(N-2k-2\ell-1)$$

$$= \sum_s \mathbf{c}(2s)\overline{\mathbf{c}}(2s+2k).$$

Thus

$$\sum_\ell \mathbf{c}(2\ell+1)\overline{\mathbf{c}}(2k+2\ell+1) + \sum_\ell \mathbf{d}(2\ell+1)\overline{\mathbf{d}}(2k+2\ell+1)$$

$$= \sum_n \mathbf{c}(n)\overline{\mathbf{c}}(n+2k) = \delta_{k0}$$

from (6.8).

3. odd-even

$$\sum_\ell \mathbf{d}(2\ell+1)\overline{\mathbf{d}}(2k+2\ell) = \sum_\ell \overline{\mathbf{c}}(N-2\ell-1)\mathbf{c}(N-2\ell-2k)$$

$$= -\sum_s \mathbf{c}(2s+1)\overline{\mathbf{c}}(2s+2k).$$

Q.E.D.

Corollary 11 *If the row vectors of $\mathbf{H}_t$ form an ON set, then the columns are also ON and $\mathbf{H}_t$ is unitary.*

To summarize, if the filter $\mathbf{L}$ satisfies the double shift orthogonality condition (6.8) then we can construct a filter $\mathbf{B}$ such that conditions (6.9), (6.10) and (6.6) hold. Thus $\mathbf{H}_t$ is unitary provided double shift orthogonality holds for the rows of the filter $\mathbf{L}$.

If $\mathbf{H}_t$ *is* unitary, then (6.6) shows us how to construct a synthesis filter bank to reconstruct the signal:

$$\overline{\mathbf{L}}^{\mathrm{tr}}\mathbf{L} + \overline{\mathbf{B}}^{\mathrm{tr}}\mathbf{B} = \mathbf{I}$$

Now $\mathbf{L} = (\downarrow 2)\mathbf{C}$ and $\mathbf{B} = (\downarrow 2)\mathbf{D}$. Using the fact that the transpose of the product of two matrices is the product of the transposed matrices in the reverse order,

$$(\mathbf{EF})^{\mathrm{tr}} = \mathbf{F}^{\mathrm{tr}}\mathbf{E}^{\mathrm{tr}},$$

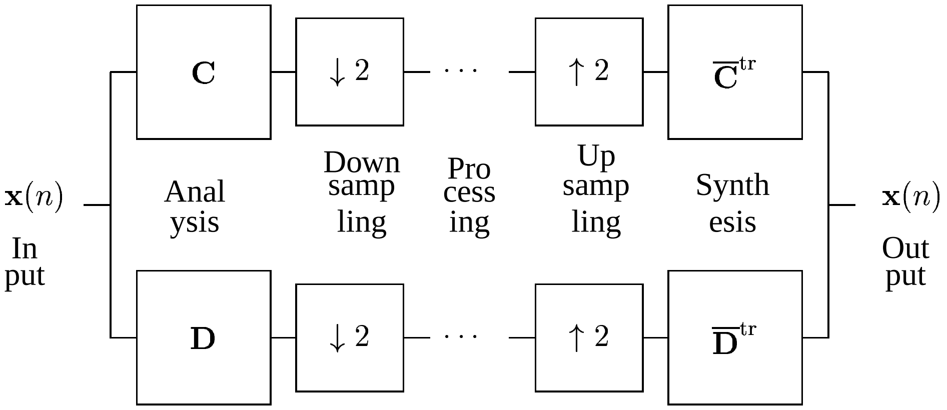

Figure 6.15: Analysis-Processing-Synthesis 2-channel filter bank system

and that $(\downarrow 2)^{\mathrm{tr}} = (\uparrow 2)$, see (6.4),(6.5), we have

$$\overline{\mathbf{L}}^{\mathrm{tr}} = \overline{\mathbf{C}}^{\mathrm{tr}}(\downarrow 2)^{\mathrm{tr}} = \overline{\mathbf{C}}^{\mathrm{tr}}(\uparrow 2),$$

$$\overline{\mathbf{B}}^{\mathrm{tr}} = \overline{\mathbf{D}}^{\mathrm{tr}}(\downarrow 2)^{\mathrm{tr}} = \overline{\mathbf{D}}^{\mathrm{tr}}(\uparrow 2).$$

Now, remembering that the order in which we apply operators in (6.6) is from right to left, we see that we have the picture of Figure 6.15.

We attach each channel of our two filter bank analysis system to a channel of a two filter bank synthesis system. On the upper channel the analysis filter $\mathbf{C}$ is applied, followed by downsampling. The output is first upsampled by the upper channel of the synthesis filter bank (which inserts zeros between successive terms of the upper analysis filter) and then filtered by $\overline{\mathbf{C}}^{\mathrm{tr}}$. On the lower channel the analysis filter $\mathbf{D}$ is applied, followed by downsampling. The output is first upsampled by the lower channel of the synthesis filter bank and then filtered by $\overline{\mathbf{D}}^{\mathrm{tr}}$. The outputs of the two channels of the synthesis filter bank are then added to reproduce the original signal.

There is still one problem. The transpose conjugate looks like Figure 6.16. This filter is *not* causal! The output of the filter at time t depends on the input at times $t+k$, $k = 0, 1, \cdots, N$. To ensure that we have causal filters we insert time delays $\mathbf{S}^N$ before the action of the synthesis filters, i.e., we replace $\overline{\mathbf{C}}^{\mathrm{tr}}$ by $\overline{\mathbf{C}}^{\mathrm{tr}}\mathbf{S}^N$

$$\overline{\mathbf{C}}^{\rm tr} = \begin{bmatrix} \cdot & \cdot & & & \cdot & \cdot \\ \cdot & \overline{\mathbf{c}}(0) & \overline{\mathbf{c}}(1) & \overline{\mathbf{c}}(2) & \overline{\mathbf{c}}(3) & \cdot \\ \cdot & 0 & \overline{\mathbf{c}}(0) & \overline{\mathbf{c}}(1) & \overline{\mathbf{c}}(2) & \cdot \\ \cdot & 0 & 0 & \overline{\mathbf{c}}(0) & \overline{\mathbf{c}}(1) & \cdot \\ \cdot & 0 & 0 & 0 & \overline{\mathbf{c}}(0) & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \end{bmatrix}.$$

Figure 6.16: $\overline{\mathbf{C}}^{\rm tr}$ matrix

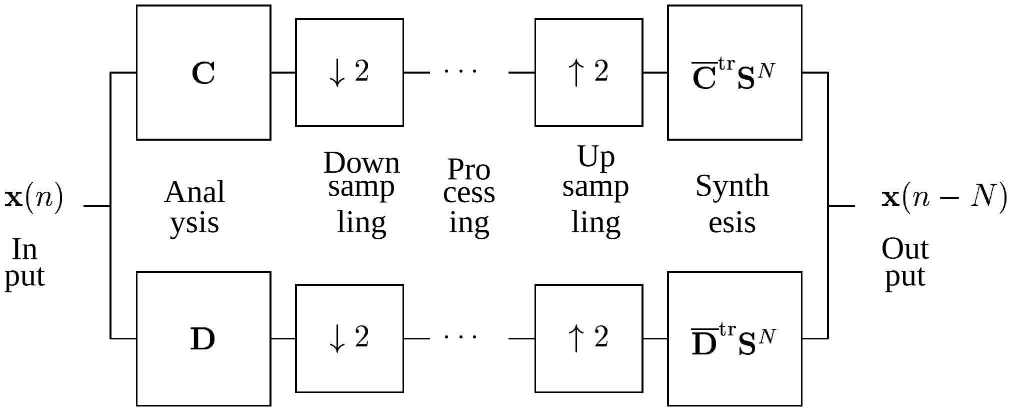

Figure 6.17: Causal 2-channel filter bank system

and $\overline{\mathbf{D}}^{\rm tr}$ by $\overline{\mathbf{D}}^{\rm tr}\mathbf{S}^N$. The resulting filters are causal, and we have reproduced the original signal with a time delay of N, see Figure 6.17.

Are there filters that actually satisfy these conditions? In the next section we will exhibit a simple solution for $N = 1$. The derivation of solutions for $N = 3, 5, \cdots$ is highly nontrivial but highly interesting, as we shall see.

6.6 A perfect reconstruction filter bank with $N = 1$

¿From the results of the last section, we can design a two-channel filter bank $\mathbf{H}_t$ with perfect reconstruction provided the rows of the filter $\mathbf{B}$ are double-shift orthogonal. For general N this is a strong restriction, for $N = 1$ it is satisfied by *all* filters. Since there are only two nonzero terms in a row $\mathbf{c}(1), \mathbf{c}(0)$, all double shifts of the row are automatically orthogonal to the original row vector. It is conventional to choose $\mathbf{H}_0$ to be a low pass filter, so that in the frequency domain $H_0(0) = 1, H_0(\pi) = 0$.

This uniquely determines $\mathbf{H}_0$. It is the moving average $\mathbf{H}_0 = \frac{1}{2}\mathbf{I} + \frac{1}{2}\mathbf{S}$. The frequency response is $H_0(\omega) = \frac{1}{2} + \frac{1}{2}e^{-in\omega}$ and the z-transform is $H_0(z) = \frac{1}{2} + \frac{1}{2}z^{-1}$. The matrix form of the action in the time domain is

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}(-1) \\ \mathbf{y}(0) \\ \mathbf{y}(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdot & \cdot \\ \cdot & \cdot & \frac{1}{2} & \frac{1}{2} & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & \frac{1}{2} & \frac{1}{2} & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

The norm of the impulse response vector $(\frac{1}{2}, \frac{1}{2})$ is $||\mathbf{h}_0|| = \frac{1}{\sqrt{2}}$. Applying the conjugate alternating flip to $\mathbf{h}_0$ we get the impulse response function $(\frac{1}{2}, -\frac{1}{2})$ of the moving difference filter, a high pass filter. Thus $\mathbf{H}_1 = \frac{1}{2}\mathbf{I} - \frac{1}{2}\mathbf{S}$ and the matrix form of the action in the time domain is

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}(-1) \\ \mathbf{y}(0) \\ \mathbf{y}(1) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & -\frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdot & \cdot \\ \cdot & \cdot & -\frac{1}{2} & \frac{1}{2} & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & -\frac{1}{2} & \frac{1}{2} & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

The $\mathbf{L}$ filter action looks like

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}_0(-2) \\ \mathbf{y}_0(0) \\ \mathbf{y}_0(2) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdot & \cdot \\ \cdot & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix},$$

$$\mathbf{H}_t = \begin{bmatrix} \mathbf{L} \\ \mathbf{B} \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdot & \cdot \\ \cdot & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdot & \cdot \\ \cdot & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix}.$$

Figure 6.18: Analysis filter bank

and the $\mathbf{B}$ filter action looks like

$$\begin{bmatrix} \cdot \\ \cdot \\ \mathbf{y}_1(-2) \\ \mathbf{y}_1(0) \\ \mathbf{y}_1(2) \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdot & \cdot \\ \cdot & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{x}(-2) \\ \mathbf{x}(-1) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \cdot \\ \cdot \end{bmatrix}.$$

The analysis part of the filter bank is pictured in Figure 6.18.

The synthesis part of the filter bank is pictured in Figure 6.19. The outputs of the upper and lower channels of the analysis filter bank are

$$(\downarrow 2)\mathbf{C}\mathbf{x}(n) = \frac{1}{\sqrt{2}}(\mathbf{x}(2n)+\mathbf{x}(2n-1)), \qquad (\downarrow 2)\mathbf{D}\mathbf{x}(n) = \frac{1}{\sqrt{2}}(\mathbf{x}(2n)-\mathbf{x}(2n-1)),$$

and we see that full information about the signal is still present. The result of upsampling the outputs of the analysis filters is

$$(\uparrow 2)(\downarrow 2)\mathbf{C}\mathbf{x}(2n+1) = 0, \qquad (\uparrow 2)(\downarrow 2)\mathbf{C}\mathbf{x}(2n) = \frac{1}{\sqrt{2}}(\mathbf{x}(2n) + \mathbf{x}(2n-1))$$

and

$$(\uparrow 2)(\downarrow 2)\mathbf{D}\mathbf{x}(2n+1) = 0, \qquad (\uparrow 2)(\downarrow 2)\mathbf{D}\mathbf{x}(2n) = \frac{1}{\sqrt{2}}(\mathbf{x}(2n) - \mathbf{x}(2n-1)).$$

$$\overline{\mathbf{H}}_t^{\mathrm{tr}} = \left[\overline{\mathbf{L}}^{\mathrm{tr}} \quad \overline{\mathbf{B}}^{\mathrm{tr}}\right] = \begin{bmatrix} \cdot & & & & \cdot & \cdot & & & & \cdot \\ \cdot & \frac{1}{\sqrt{2}} & 0 & 0 & \cdot & \cdot & \frac{1}{\sqrt{2}} & 0 & 0 & \cdot \\ \cdot & 0 & \frac{1}{\sqrt{2}} & 0 & \cdot & \cdot & 0 & -\frac{1}{\sqrt{2}} & 0 & \cdot \\ \cdot & 0 & \frac{1}{\sqrt{2}} & 0 & \cdot & \cdot & 0 & \frac{1}{\sqrt{2}} & 0 & \cdot \\ \cdot & 0 & 0 & \frac{1}{\sqrt{2}} & \cdot & \cdot & 0 & 0 & -\frac{1}{\sqrt{2}} & \cdot \\ \cdot & 0 & 0 & \frac{1}{\sqrt{2}} & \cdot & \cdot & 0 & 0 & \frac{1}{\sqrt{2}} & \cdot \\ \cdot & & & & \cdot & \cdot & & & & \cdot \end{bmatrix}.$$

Figure 6.19: Synthesis filter bank

The output of the upper synthesis filter is

$$\mathbf{C}^{\mathrm{tr}}(\uparrow 2)(\downarrow 2)\mathbf{C}\mathbf{x}(2n+1) = \frac{1}{2}(\mathbf{x}(2n+2) + \mathbf{x}(2n+1)),$$

$$\mathbf{C}^{\mathrm{tr}}(\uparrow 2)(\downarrow 2)\mathbf{C}\mathbf{x}(2n) = \frac{1}{2}(\mathbf{x}(2n) + \mathbf{x}(2n-1))$$

and the output of the lower synthesis filter is

$$\mathbf{D}^{\mathrm{tr}}(\uparrow 2)(\downarrow 2)\mathbf{D}\mathbf{x}(2n+1) = \frac{1}{2}(-\mathbf{x}(2n+2) + \mathbf{x}(2n+1)),$$

$$\mathbf{D}^{\mathrm{tr}}(\uparrow 2)(\downarrow 2)\mathbf{D}\mathbf{x}(2n) = \frac{1}{2}(\mathbf{x}(2n) - \mathbf{x}(2n-1)).$$

Delaying each filter by 1 unit for causality and then adding the outputs of the two filters we get at the nth step $\mathbf{x}(n-1)$, the original signal with a delay of 1.

6.7 Perfect reconstruction for two-channel filter banks. The view from the frequency domain.

The constructions of the preceding two sections can be clarified and generalized by examining them in the frequency domain. The filter action of convolution or multiplication by an infinite Toeplitz matrix in the time domain is replaced by multiplication by the Fourier transform or the z-transform in the frequency domain.

Let's first examine the unitarity conditions of section 6.5. Denote the Fourier transform of the impulse response vector $\mathbf{c}$ of the filter $\mathbf{C} = \frac{1}{||\mathbf{h}||}\mathbf{H}_0$ by $C(\omega)$.

Then the orthonormality of the (double-shifted) rows of $\mathbf{L} = (\downarrow 2)\mathbf{C}$ is

$$\int_{-\pi}^{\pi} e^{2ik\omega}|C(\omega)|^2 d\omega = 2\pi\delta_{k0}, \tag{6.15}$$

for integer k. Since $C(\omega) = \sum_{n=0}^{N} \mathbf{c}(n)e^{-in\omega}$, this means that the expansion of $|C(\omega)|^2$ looks like

$$|C(\omega)|^2 = 1 + \sum_{m=1}^{N} \left(a_m \cos(2m-1)\omega + b_m \sin(2m-1)\omega\right),$$

i.e., no nonzero even powers of $e^{i\omega}$ occur in the expansion. For $N = 1$ this condition is identically satisfied. For $N = 3, 5, \cdots$ it is very restrictive. An equivalent but more compact way of expressing the double-shift orthogonality in the frequency domain is

$$|C(\omega)|^2 + |C(\omega+\pi)|^2 = 2. \tag{6.16}$$

Denote the Fourier transform of the impulse response vector $\mathbf{d}$ of the filter $\mathbf{D} = \frac{1}{||\mathbf{h_1}||}\mathbf{H}_1$ by $D(\omega)$. Then the orthogonality of the (double-shifted) rows of $\mathbf{B} = (\downarrow 2)\mathbf{D}$ to the rows of $\mathbf{L}$ is expressed as

$$\int_{-\pi}^{\pi} e^{2ik\omega}C(\omega)\overline{D}(\omega)d\omega = 0 \tag{6.17}$$

for all integers k. If we take $\mathbf{d}$ to be the conjugate alternating flip of $\mathbf{c}$, then we have

$$D(\omega) = e^{-iN\omega}\overline{C}(\pi+\omega).$$

The condition (6.17) for the orthogonality of the rows of $\mathbf{L}$ and $\mathbf{B}$ becomes

$$\int_{-\pi}^{\pi} e^{i(2k+N)\omega}C(\omega)C(\pi+\omega)d\omega = (-1)^N \int_{-\pi}^{\pi} e^{i(2k+N)\phi}C(\pi+\phi)C(\phi)d\phi = 0,$$

where $\phi = \pi + \omega$, (since N is odd and $C(\omega)$ is 2π-periodic). Similarly, it is easy to show that double-shift orthogonality holds for the rows of $\mathbf{D}$:

$$|D(\omega)|^2 + |D(\omega+\pi)|^2 = 2. \tag{6.18}$$

A natural question to ask at this point is whether there are possibilities for the filter $\mathbf{d}$ other than the conjugate alternating flip of $\mathbf{c}$. The answer is no! Note that the condition (6.17) is equivalent to

$$C(\omega)\overline{D}(\omega) + C(\omega+\pi)\overline{D}(\omega+\pi) = 0. \tag{6.19}$$

Theorem 36 *If the filters* $\mathbf{c}$ *and* $\mathbf{d}$ *satisfy the double-shift orthogonality conditions (6.16), (6.16), and (6.19), then there is a constant* γ *such that* $|\gamma| = 1$ *and*

$$D(\omega) = \gamma e^{-iN\omega}\overline{C}(\pi + \omega).$$

PROOF: Suppose conditions (6.16), (6.18), and (6.19) are satisfied. Then N must be an odd integer, and we choose it to be the smallest odd integer possible. Set

$$D(\omega) = Q(e^{i\omega})e^{-iN\omega}\overline{C}(\pi + \omega), \tag{6.20}$$

for some function F. Since C and D are trigonometric polynomials in $e^{-i\omega}$ of order N, it follows that we can write

$$Q(z) = \frac{F(z)}{G(z)}, \quad F(z) = \sum_{j=0}^{N} \alpha_j z^j, \quad G(z) = \sum_{j=0}^{N} \beta_j z^j.$$

where $\alpha_0\alpha_N \neq 0$, $\beta_0\beta_N \neq 0$. Substituting the expression (6.20) for D into (6.19) and using the fact that N is odd we obtain

$$Q(z) = Q(-z).$$

Substituting into (6.16) and (6.18) we further obtain

$$|Q(e^{i\omega})| = 1.$$

Thus we have the identities

$$\frac{F(z)}{G(z)} = \frac{F(-z)}{G(-z)}, \quad |\frac{F(e^{i\omega})}{G(e^{i\omega})}| = 1. \tag{6.21}$$

From the first identity (6.21) we see that if r_j is a root of the polynomial G, then so is $-r_j$. Since N is odd, the only possibility is that F and G have an odd number of roots in common and, cancelling the common factors we have

$$Q(z) = \frac{F(z)}{G(z)} = \frac{f(z)}{g(z)} = \frac{\gamma(z^2 - s_1^2)\cdots(z^2 - s_M^2)}{(z^2 - r_1^2)\cdots(z^2 - r_M^2)},$$

i.e., the polynomials f and g are relatively prime, of even order $2M$ and all of their roots occur in ± pairs. Since D and C are trigonometric polynomials, this means that $f(e^{i\omega})$ is a factor of D. If $M > 0$ then $Q(\pm s_1) = 0$ so, considered as

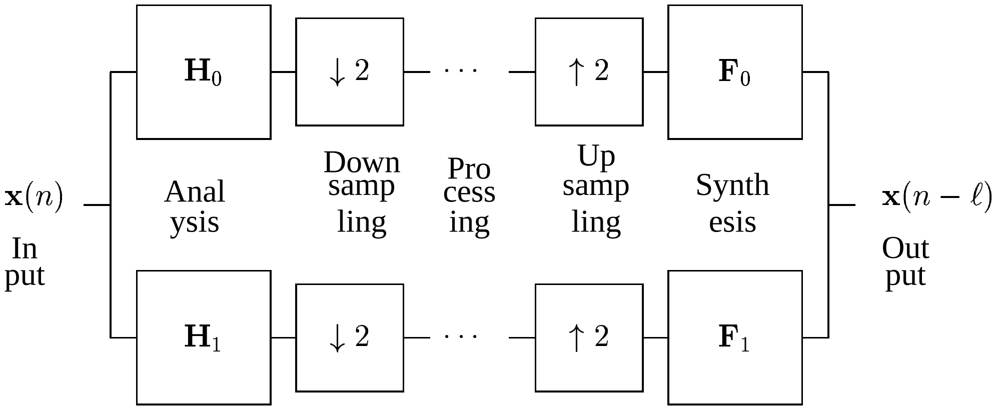

Figure 6.20: Perfect reconstruction 2-channel filter bank

a function of $e^{i\omega}$, $D(-s_1) = D(s_1) = 0$. This contradicts the condition (6.18). Hence, $M = 0$ and, from the second condition (6.21), we have $|\gamma| = 1$. Q.E.D.

If we choose $\mathbf{H}_0$ to be a low pass filter, so that $H_0(0) = 1, H_0(\pi) = 0$ then the conjugate alternating flip will have $H_1(0) = 0, |H_1(\pi)| = 1$ so that $\mathbf{H}_1$ will be a high pass filter.

Now we are ready to investigate the general conditions for perfect reconstruction for a two-channel filter bank. The picture that we have in mind is that of Figure 6.20. The analysis filter $\mathbf{H}_0$ will be low pass and the analysis filter $\mathbf{H}_1$ will be high pass. We will not impose unitarity, but the less restrictive condition of perfect reconstruction (with delay). This will require that the row and column vectors of $\mathbf{H}_t$ are *biorthogonal*. Unitarity is a special case of this.

The operator condition for perfect reconstruction with delay ℓ is

$$\mathbf{F}_0(\uparrow 2)(\downarrow 2)\mathbf{H}_0 + \mathbf{F}_1(\uparrow 2)(\downarrow 2)\mathbf{H}_1 = \mathbf{S}^\ell$$

where $\mathbf{S}$ is the shift. If we apply the operators on both sides of this requirement to a signal $\mathbf{x} = \{\mathbf{x}(n)\}$ and take the z-transform, we find

$$\frac{1}{2}F_0(z)\left[H_0(z)X(z) + H_0(-z)X(-z)\right] + \frac{1}{2}F_1(z)\left[H_1(z)X(z) + H_1(-z)X(-z)\right]$$
$$= z^{-\ell}X(z), \qquad (6.22)$$

where $X(z)$ is the z-transform of $\mathbf{x}$. The coefficient of $X(-z)$ on the left-hand side of this equation is an aliasing term, due to the downsampling and upsampling. For perfect reconstruction of a general signal $X(z)$ this coefficient must vanish. Thus we have

Theorem 37 *A 2-channel filter bank gives perfect reconstruction when*

$$\text{No distortion}: F_0(z)H_0(z) + F_1(z)H_1(z) = 2z^{-\ell} \tag{6.23}$$

$$\text{Alias cancellation}: F_0(z)H_0(-z) + F_1(z)H_1(-z) = 0 \tag{6.24}$$

In matrix form this reads

$$[F_0(z) \quad F_1(z)] \begin{bmatrix} H_0(z) & H_0(-z) \\ H_1(z) & H_1(-z) \end{bmatrix} = [2z^{-\ell} \quad 0],$$

where the 2×2 matrix is the *analysis modulation matrix* $\mathbf{H}_m(z)$.

We can solve the alias cancellation requirement (6.24) by defining the synthesis filters in terms of the analysis filters:

$$F_0(z) = H_1(-z), \qquad F_1(z) = -H_0(-z) \tag{6.25}$$

Now we focus on the no distortion requirement (6.23). We introduce the (lowpass) *product filter*

$$P_0(z) = F_0(z)H_0(z)$$

and the (high pass) product filter

$$P_1(z) = F_1(z)H_1(z).$$

¿From our solution(6.25) of the alias cancellation requirement we have $P_0(z) = H_0(z)H_1(-z)$ and $P_1(z) = -H_0(-z)H_1(z) = -P_0(-z)$. Thus the no distortion requirement reads

$$P_0(z) - P_0(-z) = 2z^{-\ell}. \tag{6.26}$$

Note that the even powers of z in $P_0(z)$ cancel out of (6.26). The restriction is only on the odd powers. This also tells us the ℓ is an odd integer. (In particular, it can never be 0.)

The construction of a perfect reconstruction 2-channel filter bank has been reduced to two steps:

1. Design the lowpass filter P_0 satisfying (6.26).

2. Factor P_0 into $F_0 H_0$, and use the alias cancellation solution to get F_1, H_1.

A further simplification involves recentering P_0 to factor out the delay term. Set $P(z) = z^{\ell} P_0(z)$. Then equation (6.26) becomes the *halfband filter* equation

$$P(z) + P(-z) = 2. \tag{6.27}$$

This equation says the coefficients of the even powers of z in $P(z)$ vanish, except for the constant term, which is 1. The coefficients of the odd powers of z are undetermined design parameters for the filter bank.

In terms of the analysis modulation matrix, and the *synthesis modulation matrix* that will be defined here, the alias cancellation and no distortion conditions read

$$\begin{bmatrix} F_0(z) & F_1(z) \\ F_0(-z) & F_1(-z) \end{bmatrix} \begin{bmatrix} H_0(z) & H_0(-z) \\ H_1(z) & H_1(-z) \end{bmatrix} = \begin{bmatrix} 2z^{-\ell} & 0 \\ o & 2(-z)^{-\ell} \end{bmatrix},$$

where the 2×2 F-matrix is the *synthesis modulation matrix* $\mathbf{F}_m(z)$. (Note the transpose distinction between $\mathbf{H}_m(z)$ and $\mathbf{F}_m(z)$.) If we recenter the filters then the matrix condition reads

$$\mathbf{F}_m(z)\mathbf{H}_m(z) = 2\mathbf{I} \tag{6.28}$$

To make contact with our earlier work on perfect reconstruction by unitarity, note that if we define $H_1(z)$ from $H_0(z)$ through the conjugate alternating flip (the condition for unitarity)

$$H_1(z) = z^{-N}\overline{H}_0(-\overline{z}^{-1}).$$

then $P_0(z) = z^{-N}\overline{H}_0(\overline{z}^{-1})H_0(z)$. Setting $z = e^{i\omega}$ and taking the complex conjugate of both sides of (6.26) we see that $N = \ell$. Thus in this case,

$$P(\omega) = \overline{H}_0(\omega)H_0(\omega) = |H_0(\omega)|^2.$$

NOTE: Any trigonometric polynomial of the form

$$P(z) = 1 + \sum_{n=-M}^{M+1} a_n z^{-(2n-1)}$$

will satisfy equation (6.27). The constants a_n are design parameters that we can adjust to achieve desired performance from the filter bank. Once $P(z)$ is chosen then we have to factor it as $P(z) = H_0(z)F_0(z)$. In theory this can always be done. Indeed $z^{2M+1}P(z)$ is a true polynomial in z and, by the fundamental theorem of algebra, polynomials over the complex numbers can always be factored completely: $z^{2M+1}P(z) = A\prod_j(z - z_j)$. Then we can define H_0 and F_0 (but not uniquely!) by assigning some of the factors to H_0 and some to F_0. If we want H_0 to be a low pass filter then we must require that $z = -1$ is a root of $P(z)$; if F_0 is also to be low pass then $P(z)$ must have -1 as a double root. If $P(z)$ is to correspond to a unitary filter bank then we must have $P(e^{i\omega}) = |H_0(e^{i\omega})|^2 \geq 0$ which is a strong restriction on the roots of $P(z)$.

6.8 Half Band Filters and Spectral Factorization

We return to our consideration of unitary 2-channel filter banks. We have reduced the design problem for these filter banks to the construction of a low pass filter $\mathbf{C}$ whose rows satisfy the double-shift orthonormality requirement. In the frequency domain this takes the form

$$|C(\omega)|^2 + |C(\omega+\pi)|^2 = 2.$$

Recall that $C(\omega) = \sum_{n=0}^{N}\mathbf{c}(n)e^{-in\omega}$. To determine the possible ways of constructing $C(\omega)$ we focus our attention on the half band filter $P(\omega) = |C(\omega)|^2$, called the *power spectral response* of $\mathbf{C}$. The frequency requirement on C can now be written as the half band filter condition (6.27)

$$P(z) + P(-z) = 2.$$

Note also that

$$P(\omega) = \sum_{n=-N}^{N}\mathbf{p}(n)e^{-in\omega} = \left(\sum_{k=0}^{N}\mathbf{c}(k)e^{-ik\omega}\right)\left(\sum_{k=0}^{N}\overline{\mathbf{c}}(k)e^{ik\omega}\right).$$

Thus

$$\mathbf{p}(n) = \sum_k \mathbf{c}(k)\overline{\mathbf{c}}(k-n) = \mathbf{c} * \overline{\mathbf{c}}^{\mathrm{T}}(n),$$

where $\mathbf{c}^{\mathrm{T}}(n) = \mathbf{c}(N-n)$ is the time reversal of $\mathbf{c}$. Since $P(\omega) \geq 0$ we have $\mathbf{p}(-n) = \overline{\mathbf{p}}(n)$. In terms of matrices we have $\mathbf{P} = \mathbf{C}\mathbf{C}^*$ where $\mathbf{C}^* = \overline{\mathbf{C}}^{\mathrm{tr}}$. ($\mathbf{P}$ is a nonnegative definite Toeplitz matrix.) The even coefficients of $\mathbf{p}$ can be obtained from the half band filter condition (6.27):

$$\mathbf{p}(2m) = \sum_k \mathbf{c}(k)\overline{\mathbf{c}}(k-2m) = \delta_{m0}, \tag{6.29}$$

i.e., $\mathbf{p}(2m) = \delta_{m0}$. The odd coefficients of $\mathbf{p}$ are undetermined. Note also that $\mathbf{P}$ is *not* a causal filter. One further comment: since $\mathbf{C}$ is a low pass filter $C(\pi) = 0$ and $C(0) = \sqrt{2}$. Thus $P(\omega) \geq 0$ for all ω, $P(0) = 2$ and $P(\pi) = 0$.

If we find a nonnegative polynomial half band filter $P(z)$, we are guaranteed that it can be factored as a perfect square.

Theorem 38 *(Fejér-Riesz Theorem) A trigonometric polynomial*

$$p(e^{-i\omega}) = \sum_{n=-N}^{N} \mathbf{p}(n) e^{-in\omega}$$

which is real and nonnegative for all ω, can be expressed in the form

$$p(e^{-i\omega}) = |C(e^{-i\omega})|^2$$

where $C(z) = \sum_{j=0}^{N} \mathbf{c}(j) z^{-j}$ is a polynomial. The polynomial $C(z)$ can be chosen such that it has no roots outside the unit disk $|z| > 1$, in which case it is unique up to multiplication by a complex constant of modulus 1.

We will prove this shortly. First some examples.

Example 4 $N = 1$

$$P(z) = 1 + \frac{z^{-1} + z}{2}, \qquad \text{or } P(\omega) = 1 + \cos\omega.$$

Here $\mathbf{p}(0) = 1, \mathbf{p}(1) = \mathbf{p}(-1) = \frac{1}{2}$. This factors as

$$P(\omega) = |C(\omega)|^2 = \frac{1}{2}(1 + e^{-i\omega})(1 + e^{i\omega}) = 1 + \cos\omega$$

and leads to the moving average filter $C(\omega)$.

Example 5 $N = 3$ *The Daubechies 4-tap filter.*

$$P(z) = (1 + \frac{z^{-1} + z}{2})^2 (1 - \frac{z^{-1} + z}{4}), \qquad \text{or } P(\omega) = (1 + \cos\omega)^2 (1 - \frac{1}{2}\cos\omega).$$

Here

$$P(z) = -\frac{1}{16} z^3 + \frac{9}{16} z + 1 + \frac{9}{16} z^{-1} - \frac{1}{16} z^{-3}.$$

Note that there are no nonzero even powers of z in $P(z)$. $P(\omega) \geq 0$ because one factor is a perfect square and the other factor $(1 - \frac{1}{2}\cos\omega) > 0$. Factoring $P(z)$ isn't trivial, but is not too hard because we have already factored the term $1 + \frac{z^{-1}+z}{2}$ in our first example. Thus we have only to factor $(1 - \frac{z^{-1}+z}{4}) = (a^+ + a^- z^{-1})(a^+ + a^- z)$. The result is $a^\pm = (1 \pm \sqrt{3})/\sqrt{8}$. Finally, we get

$$C(z) = \frac{1}{4\sqrt{2}}(1+z^{-1})^2\left((1+\sqrt{3}) + (1-\sqrt{3})z^{-1}\right)$$

$$= \frac{1}{4\sqrt{2}}\left((1+\sqrt{3}) + (3+\sqrt{3})z^{-1} + (3-\sqrt{3})z^{-2} + (1-\sqrt{3})z^{-3}\right). \quad (6.30)$$

NOTE: Only the expressions a^+a^- and $(a^+)^2 + (a^-)^2$ were determined by the above calculation. We chose the solution such that all of the roots were on or inside the circle $|z| = 1$. There are 4 possible solutions and all lead to PR filter banks, though not all to unitary filter banks. Instead of choosing the factors so that $P = |C|^2$ we can divide them in a different way to get $P = F_0 H_0$ where F_0 is not the conjugate of H_0. This would be a biorthogonal filter bank. 2) Due to the repeated factor $(1+z^{-1})^2$ in $C(z)$, it follows that $C(\omega)$ has a double zero at $\omega = \pi$. Thus $C(\pi) = 0$ and $C'(\pi) = 0$ and the response is *flat*. Similarly the response is flat at $\omega = 0$ where the derivative also vanishes. We shall see that it is highly desirable to maximize the number of derivatives of the low pass filter Fourier transform that vanish near $\omega = 0$ and $\omega = \pi$, both for filters and for application to wavelets. Note that the flatness property means that the filter has a relatively wide pass band and then a fast transition to a relatively wide stop band.

Example 6 *An ideal filter: The brick wall filter. It is easy to find solutions of the equation*

$$|C(\omega)|^2 + |C(\omega+\pi)|^2 = 2 \quad (6.31)$$

if we are not restricted to FIR filters, i.e., to trigonometric polynomials. Indeed the ideal low pass (or brick wall) filter is an obvious solution. Here

$$C(\omega) = \begin{cases} \sqrt{2}, & 0 \leq |\omega| < \frac{\pi}{2} \\ 0, & \frac{\pi}{2} \leq |\omega| \leq \pi. \end{cases}$$

The filter coefficients $\mathbf{c}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} C(\omega)e^{in\omega}d\omega$ are samples of the sinc function:

$$\mathbf{c}(n) = \frac{\sqrt{2}\sin\frac{\pi n}{2}}{\pi n} = \begin{cases} \frac{1}{\sqrt{2}}, & n = 0 \\ \pm\frac{\sqrt{2}}{\pi n}, & n \text{ odd} \\ 0, & n \text{ even}, n \neq 0. \end{cases}$$

Of course, this is an infinite impulse response filter. It satisfies double-shift orthogonality, and the companion filter is the ideal high pass filter

$$D(\omega) = \left\{ \begin{array}{ll} 0, & 0 \leq |\omega| < \frac{\pi}{2} \\ \sqrt{2}, & \frac{\pi}{2} \leq |\omega| \leq \pi. \end{array} \right.$$

This filter bank has some problems, in addition to being "ideal" and not implementable by real FIR filters. First, there is the Gibbs phenomenon that occurs at the discontinuities of the high and low pass filters. Next, this is a perfect reconstruction filter bank, but with a snag. The perfect reconstruction of the input signal follows from the Shannon sampling theorem, as the occurrence of the sinc function samples suggests. However, the Shannon sampling must occur for times ranging from $-\infty$ to ∞, so the "delay" in the perfect reconstruction is infinite!

SKETCH OF PROOF OF THE FEJÉR-RIESZ THEOREM: Since $p(e^{i\omega})$ is real we must have $\overline{p(\overline{z}^{-1})} = p(z)$. Thus, if z_j is a root of p then so is $\overline{z}_j{}^{-1}$. It follows that the roots of p that are not on the unit circle $|z| = 1$ must occur in pairs $z_j, \overline{z}_j{}^{-1}$ where $|z_j| < 1$. Since $p(e^{i\omega}) \geq 0$ each of the roots $w_k = e^{i\theta_k}, \quad 0 \leq \theta_k < 2\pi$ on the unit circle must occur with even multiplicity and the factorization must take the form

$$p(z) = \alpha^2 \Pi_{j=1}^{M} (1 - \frac{z_j}{z})(1 - z\overline{z}_j) \Pi_{k=1}^{N-M} (1 - \frac{w_k}{z})(1 - \frac{z}{w_k}) \tag{6.32}$$

where $\alpha^2 = \mathbf{p}(N)/[\Pi_j \overline{z}_j \Pi_k w_k]$ and $\alpha \geq 0$. Q.E.D.

COMMENTS ON THE PROOF:

1. If the coefficients $\mathbf{p}(n)$ of $p(z)$ are also real, as is the case with must of the examples in the text, then we can say more. We know that the roots of equations with real coefficients occur in complex conjugate pairs. Thus, if z_j is a root inside the unit circle, then so is $\overline{z}_j$, and then $z_j^{-1}, \overline{z}_j{}^{-1}$ are roots outside the unit circle. Except for the special case when z_j is real, these roots will come four at a time. Furthermore, if w_k is a root on the unit circle, then so is $\overline{w}_k$, so non real roots on the unit circle also come four at a time: $w_k, w_k, \overline{w}_k, \overline{w}_k$. The roots ± 1 if they occur, will have even multiplicity.

2. From (6.32) we can set

$$C(z) = \alpha \Pi_{j=1}^{M} (1 - \frac{z_j}{z}) \Pi_{k=1}^{N-M} (1 - \frac{w_k}{z})$$

thus uniquely defining C by the requirement that it has no roots outside the unit circle. Then $P(z) = |C(z)|^2$. On the other hand, we could factor P in different ways to get $P(z) = F(x)H(z)$. The allowable assignments of roots in the factorizations depends on the required properties of the filters F, H. For example if we want F, H to be filters with real coefficients then each complex root z_0 must be assigned to the same factor as $\overline{z}_0$.

Your text discusses a number of ways to determine the factorization in practice.

Definition 29 *An FIR filter* $\mathbf{C}$ *with impulse vector* $\mathbf{c}(n)$*,* $n = 0, 1, \cdots N$ *is self-adjoint if* $\mathbf{c}(n) = \overline{\mathbf{c}}(N - n)$*.*

This symmetry property would be a useful simplification in filter design. Note that for a filter with a real impulse vector, it means that the impulse vector is symmetric with respect to a flip about position $N/2$. The low pass filter $(1/2, 1/2)$ for $N = 1$ satisfies this requirement. Unfortunately, the following result holds:

Theorem 39 *If* $C(z)$ *is a self-adjoint unitary FIR filter, then it can have only two nonzero coefficients.*

PROOF: The self-adjoint property means $z^N C(z) = \overline{C}(\overline{z}^{-1})$. Thus if z_j is a root of $C(z)$ then so is $\overline{z}_j^{-1}$. This implies that $P(z) = \overline{C}(\overline{z}^{-1})C(z)$ has a double root at z_j. Hence all roots of the $2N$th order polynomial $z^N P(z)$ have even multiplicity and the polynomial is a perfect square

$$z^N P(z) = R(z)^2 = [\mathbf{r}(0) + \cdots \mathbf{r}(N) z^N]^2.$$

Since $P(z)$ is a half band filter, the only nonzero coefficient of an *odd* power of Z on the right hand side of this expression is the coefficient of z^N. It is easy to check that this is possible only if $R(z) = \mathbf{r}(0) + \mathbf{r}(N)z^N$, i.e., only if $R(z)$ has precisely two nonzero terms. Q.E.D.

6.9 Maxflat (Daubechies) filters

These are unitary FIR filters $\mathbf{C}$ with maximum flatness at $\omega = 0$ and $\omega = \pi$. $C(\omega)$ has exactly p zeros at $\omega = \pi$ and $N = 2p - 1$. The first member of the family, $p = 1$, is the moving average filter $(\mathbf{c}(0), \mathbf{c}(1)) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, where

$C(z) = \frac{1}{\sqrt{2}}(1+z^{-1})$. For general p the associated half band filter $P(\omega) = |C(\omega)|^2$ takes the form

$$P(z) = (\frac{1+z^{-1}}{2})^{2p}Q_{2p-2}(z), \qquad (6.33)$$

where P has degree $2N = 4p-2$. (Note that it must have $2p$ zeros at $z = -1$.) The problem is to compute $Q_{2p-2}(z)$, where the subscript denotes that Q has exactly $2p-2$ roots.

COMMENT: Since for $z = e^{i\omega}$ we have

$$(\frac{1+z^{-1}}{2})^2 = e^{-i\omega}\cos^2(\frac{\omega}{2}) = e^{-i\omega}(\frac{1+\cos\omega}{2}).$$

This means that that $P(\omega)$ has the factor $(\frac{1+\cos\omega}{2})^p$.

The condition that $C(\omega)$ has a zero of order p at $\omega = \pi$ can be expressed as

$$C(\pi) = C'(\pi) = \cdots = C^{(p-1)}(\pi) = 0. \qquad (6.34)$$

recalling that $C(\omega) = \sum_{n=0}^{2p-1} \mathbf{c}(n)e^{-in\omega}$, we see that these conditions can be expressed in the time domain as

$$\sum_{n=0}^{2p-1}(-1)^n n^k \mathbf{c}(n) = 0 \qquad k = 0, 1, \cdots, p-1. \qquad (6.35)$$

In particular, for k=0 this says

$$\sum_{\text{odd } n} \mathbf{c}(n) = \sum_{\text{even } n} \mathbf{c}(n),$$

so that the sum of the odd numbered coefficients is the same as the sum of the even numbered coefficients. We have already taken condition (6.34) into account in the expression (6.33) for $P(z)$, by requiring the P has $2p$ zeros at $z = -1$, and for $P(\omega)$ by requiring that it admits the factor $(\frac{1+\cos\omega}{2})^p$.

COMMENT: Let's look at the maximum flatness requirement at $\omega = 0$. Since $P(\omega) = |C(\omega)|^2$ and $P(0) = 2$, we can normalize $\mathbf{C}$ by the requirement $C(0) = \sqrt{2}$. Since $|C(\omega)|^2 = 2 - |C(\omega+\pi)|^2$ the flatness conditions on C at $\omega = \pi$ imply a similar vanishing of derivatives of $|C(\omega)|$ at $\omega = 0$.

We consider only the case where

$$P(\omega) = \sum_{n=1-2p}^{2p-1} \mathbf{p}(n)e^{-in\omega}$$

and the $\mathbf{p}(n)$ are *real* coefficients, i.e., the filter coefficients $\mathbf{c}(j)$ are real. Since $\mathbf{p}(0)=1$, $\mathbf{p}(2)=\mathbf{p}(4)=\cdots=\mathbf{p}(2p-2)=0$ and $\mathbf{p}(n)=\mathbf{p}(-n)$ is real for n odd, it follows that

$$P(\omega)=(\frac{1+\cos\omega}{2})^p\,\tilde{Q}_{p-1}(\cos\omega)$$

where $\tilde{Q}_{p-1}$ is a polynomial in $\cos\omega$ of order $p-1$.

REMARK: Indeed $P(\omega)$ is a linear combination of terms in $\cos n\omega$ for n odd. For any nonnegative integer n one can express $\cos n\omega$ as a polynomial of order n in $\cos\omega$. An easy way to see that this is true is to use the formula

$$e^{in\omega}=\cos n\omega+i\sin n\omega=(e^{i\omega})^n=(\cos\omega+i\sin\omega)^n\,.$$

Taking the real part of these expressions and using the binomial theorem, we obtain

$$\cos n\omega=\sum_{j=0,\cdots[\frac{n}{2}]}\left(\begin{array}{c}n\\2j\end{array}\right)(-1)^j\sin^{2j}\omega\cos^{n-2j}\omega.$$

Since $\sin^{2j}\omega=(1-\cos^2\omega)^j$, the right-hand side of the last expression is a polynomial in $\cos\omega$ of order n. Q.E.D.

We already have enough information to determine $\tilde{Q}_{p-1}(\cos\omega)$ uniquely! For convenience we introduce a new variable

$$y=\frac{1-\cos\omega}{2}\qquad\text{so that } 1-y=\frac{1+\cos\omega}{2}.$$

As ω runs over the interval $0\le\omega\le\pi$, y runs over the interval $0\le y\le 1$. Considered as a function of y, P will be a polynomial of order $2p-1$ and of the form

$$P[y]=2(1-y)^pB_p[y]$$

where B_p is a polynomial in y of order $p-1$. Furthermore $P[0]=2$. The half band filter condition now reads

$$P[y]+P[1-y]=2. \tag{6.36}$$

Thus we have

$$(1-y)^pB_p[y]=1-y^pB_p[1-y]. \tag{6.37}$$

Dividing both sides of this equation by $(1-y)^p$ we have

$$B_p[y]=(1-y)^{-p}-y^p(1-y)^{-p}B_p[1-y].$$

Since the left hand side of this identity is a polynomial in y of order $p-1$, the right hand side must also be a polynomial. Thus we can expand both terms on the right hand side in a power series in y and throw away all terms of order y^p or greater, since they must cancel to zero. Since all terms in the expansion of $y^p(1-y)^{-p}B_p[1-y]$ will be of order y^p or greater we can forget about those terms. The power series expansion of the first term is

$$(1-y)^{-p} = \sum_{k=0}^{\infty} \left(\begin{array}{c} p+k-1 \\ k \end{array} \right) y^k,$$

and taking the terms up to order y^{p-1} we find

$$B_p[y] = \sum_{k=0}^{p-1} \left(\begin{array}{c} p+k-1 \\ k \end{array} \right) y^k.$$

Theorem 40 *The only possible half band response for the maxflat filter with p zeros is*

$$P(\omega) = 2(\frac{1+\cos\omega}{2})^p \sum_{k=0}^{p-1} \left(\begin{array}{c} p+k-1 \\ k \end{array} \right) (\frac{1-\cos\omega}{2})^k. \qquad (6.38)$$

Note that $P(\omega) \geq 0$ for the maxflat filter, so that it leads to a unitary filterbank.

Strictly speaking, we have shown that the maxflat filters are uniquely determined by the expression above (the Daubechies construction), but we haven't verified that this expression actually solves the half band filter condition. Rather than verify this directly, we can use an approach due to Meyer that shows existence of a solution and gives alternate expressions for $P(\omega)$. Differentiating the expressions

$$P[y] = (1-y)^p B_p[y] = 1 - y^p B_p[1-y]$$

with respect to y, we see that $P'[y]$ is divisible by y^{p-1} and also by $(1-y)^{p-1}$. Since $P'[y]$ is a polynomial of order $2p-2$ it follows that

$$P'[y] = \kappa y^{p-1}(1-y)^{p-1},$$

for some constant κ. Differentiating the half band condition (6.36) with respect to y we get the condition

$$P'[y] - P'[1-y] = 0, \qquad (6.39)$$

which is satisfied by our explicit expression. Conversely, if $s[y] \equiv P'[y]$ satisfies (6.39) then so does $\kappa s[y]$ satisfy it and any integral $S[y]$ of this function satisfies

$$S[y] + S[1-y] = c$$

for some constant c independent of y. Thus to solve the half band filter condition we need only integrate $s[y]$ to get

$$S[y] = \kappa \int_0^y s[\tau]\, d\tau + 2$$

so that $S[0] = 2$, and choose the constant κ so that $c = 2$. Alternatively, we could compute the indefinite integral of $s[y]$ and then choose κ and the integration constant to satisfy the low pass half band filter conditions. In our case we have

$$P'[y] = \kappa y^{p-1}(1-y)^{p-1}.$$

Integrating by parts $p-1$ times (i.e., repeatedly integrating $(1-y)^n$ and differentiating y^m), we obtain

$$P[y] = -\kappa \sum_{h=0}^{p-1} \frac{(1-y)^{p+h} y^{p-1-h} [(p-1)!]^2}{(p+h)!(p-h-1)!}$$

where the integration constant has been chosen so that $P[1] = 0$. To get $c = 2$ we require

$$P[0] = -\kappa \frac{[(p-1)!]^2}{(2p-1)!} = 2$$

or $\kappa = -2(2p-1)!/[(p-1)!]^2$, so that

$$P[y] = 2(1-y)^p \sum_{h=0}^{p-1} \left(\begin{array}{c} 2p-1 \\ p+h \end{array} \right) (1-y)^h y^{p-1-h}.$$

Thus a solution exists.

Another interesting form of the solution can be obtained by changing variables from y to ω. We have $dy = \frac{1}{2}\sin\omega\, d\omega$ so

$$P'(\omega) = \frac{dy}{d\omega} P'[y] = \frac{\kappa \sin\omega}{2^{2p-1}} \sin^{2p-2}\omega = -a \sin^{2p-1}\omega.$$

Then

$$P(\omega) = 2 - a \int_0^\omega \sin^{2p-1}\omega\, d\omega \tag{6.40}$$

where the constant a is determined by the requirement $P(\pi) = 0$. Integration by parts yields

$$\int_0^\pi \sin^{2p-1}\omega\, d\omega = \frac{2^{2p-1}[(p-1)!]^2}{(2p-1)!} = \frac{\sqrt{\pi}\Gamma(p)}{\Gamma(p+\frac{1}{2})},$$

where $\Gamma(z)$ is the gamma function. Thus

$$a = \frac{2\Gamma(p+\frac{1}{2})}{\sqrt{\pi}\Gamma(p)}.$$

Stirling's formula says

$$\Gamma(z) \sim z^{z-\frac{1}{2}} e^{-z} \sqrt{2\pi} \left(1 + O(\frac{1}{z})\right),$$

so $a \sim \sqrt{4p/\pi}$ as $p \to \infty$. Since $P'(\pi/2) = -a$, we see that the slope at the center of the maxflat filter is proportional to $\sqrt{N}$. Moreover $P(\omega)$ is monotonically decreasing as ω goes from 0 to π. One can show that the transition band gets more and more narrow. Indeed, the transition from $P(\omega) = 0.98$ to 0.02 takes place over an interval of length $\frac{4}{\sqrt{N}}$.

When translated back to the z-transform, the maxflat half band filters with $2p$ zeros at $\omega = \pi$ factor to the unitary low pass Daubechies filters C with $N = 2p-1$. The notation for the Daubechies filter with $N = 2p-1$ is D_{N+1}. We have already exhibited D_4 as an example. Of course, D_2 is the "moving average" filter.

Exercise 1 *Show that it is not possible to satisfy the half band filter conditions for $P[y]$ with p zeros where $2p > N + 1$. That is, show that the number of zeros for a maxflat filter is indeed a maximum.*

Exercise 2 *Show that each of the following expressions leads to a formal solution $P[y]$ of the half band low pass filter conditions*

$$P[y] + P[1-y] = 2, \quad P[1] = 0.$$

Determine if each defines a filter. A unitary filter? A maxflat filter?

a.

$$P'[y] = \kappa$$

b.

$$P'[y] = \kappa y(1-y)(y-\frac{1}{2})^2$$

c.

$$P'[y] = \kappa y(1-y)(y-\frac{1}{2})$$

d.

$$P'[y] = \kappa y(1-y)(y-2)(y+1)$$

e.

$$P'[y] = \kappa y(1-y)(y-a)(y+a-1)$$

Chapter 7

Multiresolution Analysis

7.1 Haar wavelets

The simplest wavelets are the Haar wavelets. They were studied by Haar more than 50 years before wavelet theory came into vogue. The connection between filters and wavelets was also recognized only rather recently. We will make it apparent from the beginning. We start with the *father wavelet* or *scaling function*. For the Haar wavelets the scaling function is the *box function*

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

We can use this function and its integer translates to construct the space V_0 of all step functions of the form

$$s(t) = a_k \qquad \text{for } k \le t < k+1,$$

where the a_k are complex numbers such that $\sum_{k=-\infty}^{\infty} |a_k|^2 < \infty$. Thus $s \in V_0 \subset L^2[-\infty, \infty]$ if and only if

$$s(t) = \sum_k a_k \phi(t-k), \qquad \sum_{k=-\infty}^{\infty} |a_k|^2 < \infty.$$

Note that the $\{\phi(t-k) : \quad k = 0, \pm 1, \cdots\}$ form an ON basis for V_0. Also, the area under the father wavelet is 1:

$$\int_{-\infty}^{\infty} \phi(t) dt = 1.$$

We can approximate signals $f(t) \in L^2[-\infty, \infty]$ by projecting them on V_0 and then expanding the projection in terms of the translated scaling functions. Of course this would be a very crude approximation. To get more accuracy we can change the scale by a factor of 2.

Consider the functions $\phi(2t - k)$. They form a basis for the space V_1 of all step functions of the form

$$s(t) = a_k \qquad \text{for } \frac{k}{2} \le t < \frac{k+1}{2},$$

where $\sum_{k=-\infty}^{\infty} |a_k|^2 < \infty$. This is a larger space than V_0 because the intervals on which the step functions are constant are just $1/2$ the width of those for V_0. The functions $\{2^{1/2}\phi(2t-k) : \quad k = 0, \pm 1, \cdots\}$ form an ON basis for V_1.The scaling function also belongs to V_1. Indeed we can expand it in terms of the basis as

$$\phi(t) = \phi(2t) + \phi(2t - 1). \tag{7.2}$$

NOTE: In the next section we will study many new scaling functions ϕ. We will always require that these functions satisfy the *dilation equation*

$$\phi(t) = \sqrt{2} \sum_{k=0}^{N} \mathbf{c}(k)\phi(2t - k), \tag{7.3}$$

or, equivalently,

$$\phi(t) = 2 \sum_{k=0}^{N} \mathbf{h}(k)\phi(2t - k). \tag{7.4}$$

For the Haar scaling function $N = 1$ and $(\mathbf{h}(0), \mathbf{h}(1)) = (\frac{1}{2}, \frac{1}{2})$. From (7.4) we can easily prove

Lemma 35 *If the scaling function is normalized so that*

$$\int_{-\infty}^{\infty} \phi(t)dt = 1,$$

then $\sum_{k=0}^{N} \mathbf{h}(k) = 1$.

Returning to Haar wavelets, we can continue this rescaling procedure and define the space V_j of step functions at level j to be the Hilbert space spanned by the linear combinations of the functions $\phi(2^j t - k), \quad k = 0, \pm 1, \cdots$. These functions will be piecewise constant with discontinuities contained in the set

$$\{t = \frac{n}{2^j}, \qquad n = 0, \pm 1, \pm 2, \cdots\}.$$

The functions

$$\phi_{jk}(t) = 2^{\frac{j}{2}}\phi(2^j t - k), \qquad k = 0, \pm 1, \pm 2, \cdots$$

form an ON basis for V_j. Further we have

$$V_0 \subset V_1 \subset \cdots \subset V_{j-1} \subset V_j \subset V_{j+1} \subset \cdots$$

and the containment is strict. (Each V_j contains functions that are not in V_{j-1}.) Also, note that the dilation equation (7.2) implies that

$$\phi_{jk}(t) = \frac{1}{\sqrt{2}}[\phi_{j+1,2k}(t) + \phi_{j+1,2k+1}(t)]. \tag{7.5}$$

NOTE: Our definition of the space V_j and functions $\phi_{jk}(t)$ also makes sense for negative integers j. Thus we have

$$\cdots V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots .$$

Here is an easy way to decide in which class a step function $s(t)$ belongs:

Lemma 36

1. $s(t) \in V_0 \Leftrightarrow s(2^j t) \in V_j$
2. $s(t) \in V_j \Leftrightarrow s(2^{-j} t) \in V_0$

PROOF: $s(t)$ is a linear combination of functions $\phi(t-k)$ if and only if $s(2^j t)$ is a linear combination of functions $\phi(2^j t - k)$. Q.E.D.

Since $V_0 \subset V_1$, it is natural to look at the orthogonal complement of V_0 in V_1, i.e., to decompose each $s \in V_1$ in the form $s = s_0 + s_1$ where $s_0 \in V_0$ and $s_1 \in V_0^{\perp}$. We write

$$V_1 = V_0 \oplus W_0,$$

where $W_0 = \{s \in V_1 : \quad (s,f) = 0 \text{ for all } f \in V_0\}$. It follows that the functions in W_0 are just those in V_1 that are orthogonal to the basis vectors $\phi(t-k)$ of V_0.

Note from the dilation equation that $\phi(t-k) = \phi(2t-2k) + \phi(2t-2k-1) = 2^{-1/2}\left(\phi_{1,2k}(t) + \phi_{1,2k+1}(t)\right)$. Thus

$$(\phi_{0k}, \phi_{1\ell}) = 2^{1/2}\int_{-\infty}^{\infty} \phi(t-k)\phi(2t-\ell)dt = \begin{cases} 2^{-1/2} & \text{if } \ell = 2k, 2k+1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$s_1(t) = \sum_k a_k \phi(2t-k) \in V_1$$

belongs to W_0 if and only if $a_{2k+1} = -a_{2k}$. Thus

$$s_1 = \sum_k a_{2k} \left[\phi(2t-2k) - \phi(2t-2k-1)\right] = \sum_k a_{2k} w(t-k)$$

where

$$w(t) = \phi(2t) - \phi(2t-1) \tag{7.6}$$

is the *Haar wavelet*, or *mother wavelet*. You can check that the wavelets $w(t-k), \qquad k = 0 \pm 1, \cdots$ form an ON basis for W_0.
NOTE: In the next section we will require that associated with the father wavelet $\phi(t)$ there be a mother wavelet $w(t)$ satisfying the *wavelet equation*

$$w(t) = \sqrt{2}\sum_{k=0}^{N} \mathbf{d}(k)\phi(2t-k), \tag{7.7}$$

or, equivalently,

$$w(t) = 2\sum_{k=0}^{N} \mathbf{h}_1(k)\phi(2t-k), \tag{7.8}$$

and such that w is orthogonal to all translations $\phi(t-k)$ of the father wavelet. For the Haar scaling function $N=1$ and $(\mathbf{h}_1(0), \mathbf{h}_1(1)) = (\frac{1}{2}, -\frac{1}{2})$.

We define functions

$$w_{jk}(t) = 2^{\frac{j}{2}} w(2^j t - k) = 2^{\frac{j}{2}}(\phi(2^{j+1}t - 2k) - \phi(2^{j+1}t - 2k - 1),$$
$$k = 0, \pm 1, \pm 2, \cdots, \quad j = 1, 2, \cdots.$$

It is easy to prove

Lemma 37 *For fixed j,*

$$(w_{jk}, w_{jk'}) = \delta_{kk'}, \qquad (\phi_{jk}, w_{jk'}) = 0 \tag{7.9}$$

where $k, k' = 0, \pm 1, \cdots$.

Other properties proved above are

$$\phi_{jk}(t) = \frac{1}{\sqrt{2}}(\phi_{j+1,2k}(t) + \phi_{j+1,2k+1}(t)),$$

$$w_{jk}(t) = \frac{1}{\sqrt{2}}(\phi_{j+1,2k}(t) - \phi_{j+1,2k+1}(t)).$$

Theorem 41 *let W_j be the orthogonal complement of V_j in V_{j+1}:*

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

The wavelets $\{w_{jk}(t) : \quad k = 0, \pm 1, \cdots\}$ form an ON basis for W_j.

PROOF: From (7.9) it follows that the wavelets $\{w_{jk}\}$ form an ON set in W_j. Suppose $s \in W_j \subset V_{j+1}$. Then

$$s(t) = \sum_k a_k \phi_{j+1,k}(t)$$

and $(s, \phi_{jn}) = 0$ for all integers n. Now

$$\phi_{jn} = \frac{1}{\sqrt{2}}(\phi_{j+1,2n}(t) + \phi_{j+1,2n+1}(t)).$$

Thus

$$0 = (s, \phi_{jn}) = \frac{1}{\sqrt{2}}[(s, \phi_{j+1,2n}) + (s, \phi_{j+1,2n+1})] = \frac{1}{\sqrt{2}}(a_{2n} + a_{2n+1}).$$

Hence

$$s(t) = \sum_k a_{2k}\sqrt{2}w_{jk}(t),$$

so the set $\{w_{jk}\}$ is an ON basis for W_j. Q.E.D.

Since $V_j \oplus W_j = V_{j+1}$ for all $j \geq 0$, we can iterate on j to get $V_{j+1} = W_j \oplus V_j = W_j \oplus W_{j-1} \oplus V_{j-1}$ and so on. Thus

$$V_{j+1} = W_j \oplus W_{j-1} \oplus \cdots \oplus W_1 \oplus W_0 \oplus V_0.$$

and any $s \in V_{j+1}$ can be written uniquely in the form

$$s = \sum_{k=0}^{j} w_k + s_0 \qquad \text{where } w_k \in W_k,\ s_0 \in V_0.$$

REMARK: Note that $(w_{jk}, w_{j'k'}) = 0$ if $j \neq j'$. Indeed, suppose $j > j'$ to be definite. Then $w_{j'k'} \in W_{j'} \subset V_{j'+1} \subseteq V_j$. Since $w_{jk} \in W_j$ it must be perpendicular to $w_{j'k'}$.

Lemma 38 $(w_{jk}, w_{j'k'}) = \delta_{jj'}\delta_{kk'}$ *for* $j, j', \pm k, \pm k' = 0, 1, \cdots$.

Theorem 42

$$L^2[-\infty,\infty] = V_0 \oplus \sum_{\ell=0}^{\infty} W_\ell = V_0 \oplus W_0 \oplus W_1 \oplus \cdots,$$

so that each $f(t) \in L^2[-\infty,\infty]$ can be written uniquely in the form

$$f = f_0 + \sum_{k=0}^{\infty} w_\ell, \qquad w_\ell \in W_\ell,\ f_0 \in V_0. \tag{7.10}$$

PROOF: based on our study of Hilbert spaces, it is sufficient to show that for any $f \in L^2[-\infty,\infty]$, given $\epsilon > 0$ we can find an integer $j(\epsilon)$ and a step function $s = \sum_k a_k \phi_{jk} \in V_j$ with a finite number of nonzero a_k and such that $||f - s|| < \epsilon$. This is easy. Since the space of step functions $S^2_{[-\infty,\infty]}$ is dense in $L^2[-\infty,\infty]$ there is a step function $s'(t) \in S^2_{[-\infty,\infty]}$, nonzero on a finite number of bounded intervals, such that $||f - s'|| < \frac{\epsilon}{2}$. Then, it is clear that by choosing j sufficiently large, we can find an $s \in V_j$ with a finite number of nonzero a_k and such that $||s' - s|| < \frac{\epsilon}{2}$. Thus $||f - s|| \le ||f - s'|| + ||s' - s|| < \epsilon$. Q.E.D.

Note that for j a negative integer we can also define spaces V_j, W_j and functions ϕ_{jk}, w_{jk} in an obvious way, so that we have

$$L^2[-\infty,\infty] = V_j \oplus \sum_{\ell=j}^{\infty} W_\ell = V_j \oplus W_j \oplus W_{j+1} \oplus \cdots, \tag{7.11}$$

even for negative j. Further we can let $j \to -\infty$ to get

Corollary 12

$$L^2[-\infty,\infty] = \sum_{\ell=-\infty}^{\infty} W_\ell = \cdots W_{-1} \oplus W_0 \oplus W_1 \oplus \cdots,$$

so that each $f(t) \in L^2[-\infty,\infty]$ can be written uniquely in the form

$$f = \sum_{\ell=-\infty}^{\infty} w_\ell, \qquad w_\ell \in W_\ell. \tag{7.12}$$

In particular, $\{w_{jk} : j,k = 0,\pm 1,\pm 2,\cdots\}$ is an ON basis for $L^2[-\infty,\infty]$.

PROOF (If you understand that very function in $L^2[-\infty, \infty]$ is determined up to its values on a set of measure zero.): We will show that $\{w_{jk}\}$ is an ON basis for $L^2[-\infty, \infty]$. The proof will be complete if we can show that the space W' spanned by all finite linear combinations of the w_{jk} is dense in $L^2[-\infty, \infty]$. This is equivalent to showing that the only $g \in L^2[-\infty, \infty]$ such that $(g, f) = 0$ for all $f \in W'$ is the zero vector $g \equiv 0$. It follows immediately from (7.11) that if $(g, w_{jk}) = 0$ for all j, k then $g \in V_\ell$ for all integers ℓ. This means that, almost everywhere, g is equal to a step function that is constant on intervals of length $2^{-\ell}$. Since we can let ℓ go to $-\infty$ we see that, almost everywhere, $g(t) = c$ where c is a constant. We can't have $c \neq 0$ for otherwise g would not be square integrable. Hence $g \equiv 0$. Q.E.D.

We have a new ON basis for $L^2[-\infty, \infty]$:

$$\{\phi_{0k}, w_{jk'} : \qquad j, \pm k, \pm k' = 0, 1, \cdots\}.$$

Let's consider the space V_j for fixed j. On one hand we have the scaling function basis

$$\{\phi_{j,k} : \qquad \pm k = 0, 1, \cdots\}.$$

Then we can expand any $f_j \in V_j$ as

$$f_j = \sum_{k=-\infty}^{\infty} a_{j,k}\phi_{j,k}. \tag{7.13}$$

On the other hand we have the wavelets basis

$$\{\phi_{j-1,k}, w_{j-1,k'} : \qquad \pm k, \pm k' = 0, 1, \cdots\}$$

associated with the direct sum decomposition

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

Using this basis we can expand any $f_j \in V_j$ as

$$f_j = \sum_{k'=-\infty}^{\infty} b_{j-1,k'}w_{j-1,k'} + \sum_{k=-\infty}^{\infty} a_{j-1,k}\phi_{j-1,k}. \tag{7.14}$$

If we substitute the relations

$$\phi_{j-1,k}(t) = \frac{1}{\sqrt{2}}(\phi_{j,2k}(t) + \phi_{j,2k+1}(t)),$$

$$w_{j-1,k}(t) = \frac{1}{\sqrt{2}}(\phi_{j,2k}(t) - \phi_{j,2k+1}(t))$$

into the expansion (7.14) and compare coefficients of $\phi_{j,\ell}$ with the expansion (7.13), we obtain the fundamental recursions

$$\text{Averages(lowpass)} \quad a_{j-1,k} = \tfrac{1}{\sqrt{2}}(a_{j,2k} + a_{j,2k+1}) \tag{7.15}$$

$$\text{Differences(highpass)} \quad b_{j-1,k} = \tfrac{1}{\sqrt{2}}(a_{j,2k} - a_{j,2k+1}). \tag{7.16}$$

These equations link the Haar wavelets with the $N = 1$ unitary filter bank. Let $\mathbf{x}(k) = a_{jk}$ be a discrete signal. The result of passing this signal through the (normalized) moving average filter $\mathbf{C}$ and then downsampling is $\mathbf{y}(k) = (\downarrow 2)\mathbf{C}^T * \mathbf{x}(k) = a_{j-1,k}$, where $a_{j-1,k}$ is given by (7.15). Similarly, the result of passing the signal through the (normalized) moving difference filter $\mathbf{D}$ and then downsampling is $\mathbf{z}(k) = (\downarrow 2)\mathbf{D}^T * \mathbf{x}(k) = b_{j-1,k}$, where $b_{j-1,k}$ is given by (7.16).

NOTE: If you compare the formulas (7.15), (7.16) with the action of the filters $\mathbf{C}$, $\mathbf{D}$, you see that the correct high pass filters differ by a time reversal. The correct analysis filters are the time reversed filters $\mathbf{D}^T$, where the impulse response vector is $\mathbf{d}^T(n) = \mathbf{d}(-n)$, and $\mathbf{C}^T$. These filters are not causal. In general, the analysis recurrence relations for wavelet coefficients will involve the corresponding acausal filters $\mathbf{C}^T$ $\mathbf{D}^T$. The synthesis filters will turn out to be $\mathbf{C}$, $\mathbf{D}$ exactly.

The picture is in Figure 7.1.

We can iterate this process by inputting the output $a_{j-1,k}$ of the high pass filter to the filter bank again to compute $a_{j-2,k}, b_{j-2,k}$, etc. At each stage we save the wavelet coefficients $b_{j'k'}$ and input the scaling coefficients $a_{j'k'}$ for further processing, see Figure 7.2. The output of the final stage is the set of scaling coefficients a_{0k}. Thus our final output is the complete set of coefficients for the wavelet expansion

$$f_j = \sum_{j'=0}^{j} \sum_{k=-\infty}^{\infty} b_{j'k} w_{j'k} + \sum_{k=-\infty}^{\infty} a_{0k}\phi_{0k},$$

based on the decomposition

$$V_{j+1} = W_j \oplus W_{j-1} \oplus \cdots \oplus W_1 \oplus W_0 \oplus V_0.$$

The synthesis recursion is :

$$\begin{aligned} a_{j,2k} &= \frac{1}{\sqrt{2}}(a_{j-1,k} + b_{j-1,k}) \\ a_{j,2k+1} &= \frac{1}{\sqrt{2}}(a_{j-1,k} - b_{j-1,k}). \end{aligned} \tag{7.17}$$

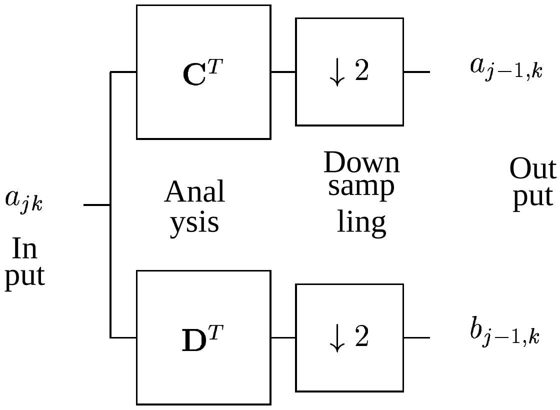

Figure 7.1: Haar Wavelet Recursion

This is exactly the output of the synthesis filter bank shown in Figure 7.3. Thus, for level j the full analysis and reconstruction picture is Figure 7.4.

COMMENTS ON HAAR WAVELETS:

1. For any $f(t) \in L^2[-\infty, \infty]$ the scaling and wavelets coefficients of f are defined by

$$\begin{aligned} a_{jk} &= (f, \phi_{jk}) = 2^{j/2}\int_{-\infty}^{\infty} f(t)\phi(2^jt-k)dt \\ &= 2^{j/2}\int_{\frac{k}{2^j}}^{\frac{k}{2^j}+\frac{1}{2^j}} f(t)dt, \qquad (7.18) \\ b_{jk} &= (f, w_{jk}) = 2^{j/2}\int_{-\infty}^{\infty} f(t)\phi(2^{j+1}t-2k)dt \\ &- 2^{j/2}\int_{-\infty}^{\infty} f(t)\phi(2^{j+1}t-2k-1)dt \\ &= 2^{j/2}\int_{\frac{k}{2^j}}^{\frac{k}{2^j}+\frac{1}{2^j}} [f(t)-f(t+\frac{1}{2^{j+1}})]dt. \qquad (7.19) \end{aligned}$$

If f is a continuous function and j is large then $a_{jk} \sim 2^{-j/2}f(\frac{k}{2^j})$. (Indeed if f has a bounded derivative we can develop an upper bound for the error of this approximation.) If f is continuously differentiable and j is large, then

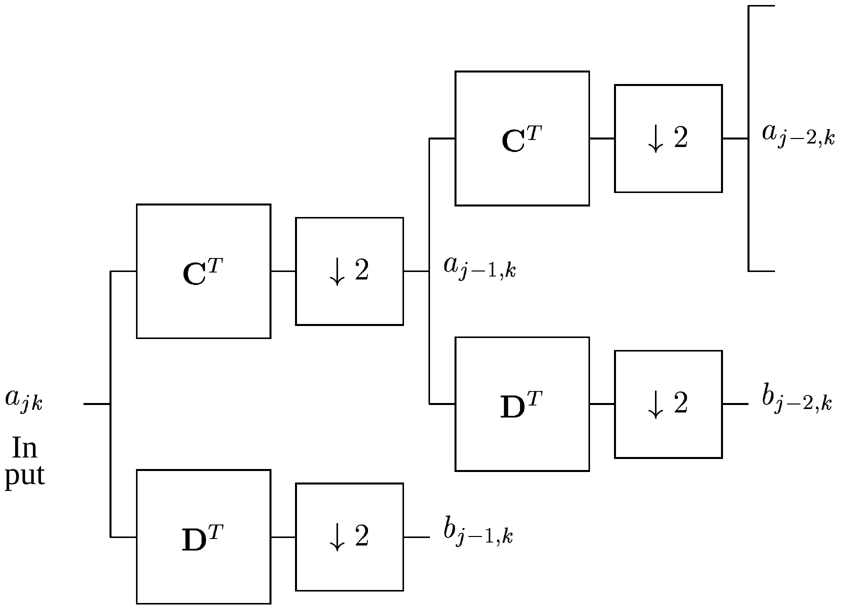

Figure 7.2: Fast Wavelet Transform

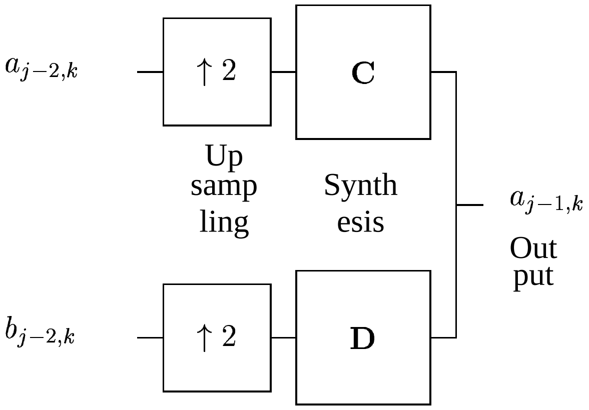

Figure 7.3: Haar wavelet inversion

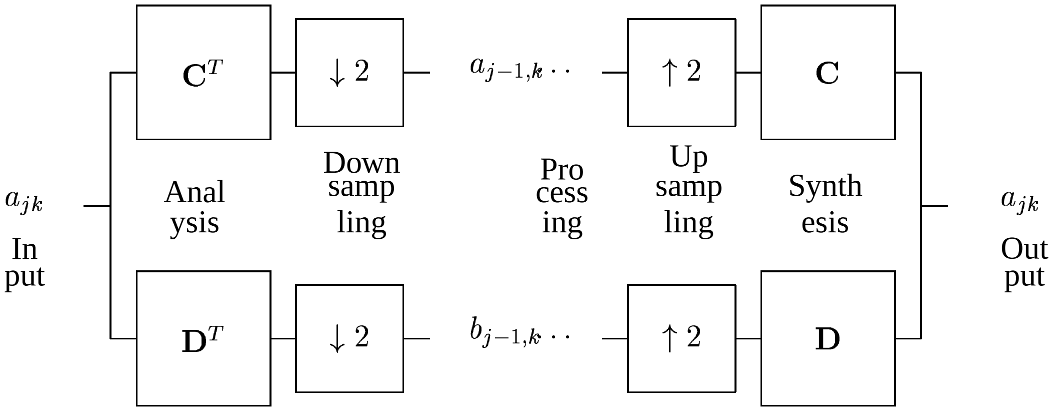

Figure 7.4: Fast Wavelet Transform and Inversion

$b_{jk} \sim -\frac{1}{2^{1+3j/2}} f'(\frac{k}{2^j})$. Again this shows that the a_{jk} capture averages of f (low pass) and the b_{jk} capture changes in f (high pass).

2. Since the scaling function $\phi(t)$ is nonzero only for $0 \le t < 1$ it follows that $\phi_{jk}(t)$ is nonzero only for $\frac{k}{2^j} \le t < \frac{k}{2^j} + \frac{1}{2j}$. Thus the coefficients a_{jk} depend only on the local behavior of $f(t)$ in that interval. Similarly for the wavelet coefficients b_{jk}. This is a dramatic difference from Fourier series or Fourier integrals where each coefficient depends on the global behavior of f. If f has compact support, then for fixed j, only a finite number of the coefficients a_{jk}, b_{jk} will be nonzero. The Haar coefficients a_{jk} enable us to track t intervals where the function becomes nonzero or large. Similarly the coefficients b_{jk} enable us to track t intervals in which f changes rapidly.

3. Given a signal f, how would we go about computing the wavelet coefficients? As a practical matter, one doesn't usually do this by evaluating the integrals (7.18) and (7.19). Suppose the signal has compact support. By translating and rescaling the time coordinate if necessary, we can assume that $f(t)$ vanishes except in the interval $[0, 1)$. Since $\phi_{jk}(t)$ is nonzero only for $\frac{k}{2^j} \le t < \frac{k}{2^j} + \frac{1}{2j}$ it follows that all of the coefficients a_{jk}, b_{jk} will vanish except when $0 \le k < 2^j$. Now suppose that f is such that for a sufficiently large integer $j = J$ we have $a_{Jk} \sim 2^{-J/2} f(\frac{k}{2^J})$. If f is differentiable we can compute how large J needs to be for a given error tolerance. We would also want to exceed the Nyquist rate. Another possibility is that f takes discrete values on the grid $t = \frac{k}{2^J}$, in which case there is no error in our assumption.

   Inputing the values $a_{Jk} = 2^{-J/2} f(\frac{k}{2^J})$ for $= 0, 1, \cdots, 2^J - 1$ we use the recursion

$$\text{Averages (lowpass)} \quad a_{j-1,k} = \frac{1}{\sqrt{2}}(a_{j,2k} + a_{j,2k+1}) \qquad (7.20)$$

$$\text{Differences (highpass)} \quad b_{j-1,k} = \frac{1}{\sqrt{2}}(a_{j,2k} - a_{j,2k+1}). \qquad (7.21)$$

   described above, see Figure 7.2, to compute the wavelet coefficients b_{jk}, $j = 0, 1, \cdots, J-1, \quad k = 0, 1, \cdots 2^j - 1$ and a_{00}.

   The input consists of 2^J numbers. The output consists of $\sum_{j=0}^{J-1} 2^j + 1 = 2^J$ numbers. The algorithm is very efficient. Each recurrence involves 2 multiplications by the factor $\frac{1}{\sqrt{2}}$. At level j there are $2 \cdot 2^j$ such recurrences. thus the total number of multiplications is $2\sum_{j=0}^{J-1} 2 \cdot 2^j = 4 \cdot 2^J - 4 < 4 \cdot 2^J$.

4. The preceding algorithm is an example of the Fast Wavelet Transform (FWT). It computes 2^J wavelet coefficients from an input of 2^J function values and

does so with a number of multiplications $\sim 2^J$. Compare this with the FFT which needs $\sim J \cdot 2^J$ multiplications from an input of 2^J function values. In theory at least, the FWT is faster. The Inverse Fast Wavelet Transform is based on (7.17). (Note, however, that the FFT and the FWT compute different things. They divide the spectral band in different ways. Hence they aren't directly comparable.)

5. The FWT discussed here is based on filters with $N+1$ taps, where $N=1$. For wavelets based on more general $N+1$ tap filters (such as the Daubechies filters) , each recursion involves $N+1$ multiplications, rather than 2. Otherwise the same analysis goes through. Thus the FWT requires $\sim 2(N+1)2^J$ multiplications.

6. What would be a practical application of Haar wavelets in signal processing? Boggess and Narcowich give an example of signals from a faulty voltmeter. The analog output from the voltmeter is usually smooth, like a sine wave. However if there is a loose connection in the voltmeter there could be sharp spikes in the output, large changes in the output, but of very limited time duration. One would like to filter this "noise" out of the signal, while retaining the underlying analog readings. If the sharp bursts are on the scale $\Delta t = \frac{1}{2^{j_0}}$ then the spikes will be identifiable as large values of $|b_{j_0k}|$ for some k. We could use Haar wavelets with $J > j_0$ to analyze the signal. Then we could process the signal to identify all terms b_{j_0k} for which $|b_{j_0k}|$ exceeds a fixed tolerance level δ and set those wavelet coefficients equal to zero. Then we could resynthesize the processed signal.

7. Haar wavelets are very simple to implement. However they are terrible at approximating continuous functions. By definition, any truncated Haar wavelet expansion is a step function. The Daubechies wavelets to come are continuous and are much better for this type of approximation.

7.2 The Multiresolution Structure

The Haar wavelets of the last section, with their associated nested subspaces that span L^2 are the simplest example of resolution analysis. We give the full definition here. It is the main structure that we shall use for the study of wavelets, though not the only one. Almost immediately we will see striking parallels with the study of filter banks.

Figure 7.5: Haar Analysis of a Signal

This is output from the Wavelet Toolbox of Matlab. The signal $s = a_0$ is sampled at $1024 = 2^{10}$ points, so $J = 10$ and s is assumed to be in the space V_{10}. The signal is taken to be zero at all points $k/2^{10}$, except for $k = 0, 1, \cdots, 2^{10} - 1$. The approximations a_ℓ (the averages) are the projections of s on the subspaces $V_{10-\ell}$ for $\ell = 1, \cdots, 6$. The lowest level approximation a_6 is the projection on the subspace V_4. There are only 16 distinct values at this lowest level. The approximations d_ℓ (the differences) are the projections of s on the wavelet subspaces $W_{10-\ell}$.

Figure 7.6: Tree Stucture of Haar Analysis

This is output from the Wavelet Toolbox of Matlab. As before the signal $s = a_0$ is sampled at $1024 = 2^{10}$ points, so $J = 10$ and s is assumed to be in the space V_{10}. The signal can be reconstructed in a variety of manners: $s = a_6 + d_6 + d_5 + d_4 + d_3 + d_2 + d_1$, or $s = a_1 + d_1$, or $s = a_2 + d_2 + d_1$, etc. Note that the signal is a Doppler waveform with noise superimposed. The lower-order difference contain information, but the differences d_1 appear to be noise. Thus one possible way of processing this signal to reduce noise and pass on the underlying information would be to set the d_1 coefficients $b_{9,k} = 0$ and reconstruct the signal from the remaining nonzero coefficients.

Figure 7.7: Separate Components in Haar Analysis

This is output from the Wavelet Toolbox of Matlab. It shows the complete decomposition of the signal into a_ℓ and d_ℓ components.

Definition 30 *Let $\{V_j : j = \cdots, -1, 0, 1, \cdots\}$ be a sequence of subspaces of $L^2[-\infty, \infty]$ and $\phi \in V_0$. This is a multiresolution analysis for $L^2[-\infty, \infty]$ provided the following conditions hold:*

1. *The subspaces are nested: $V_j \subset V_{j+1}$.*
2. *The union of the subspaces generates L^2 : $\overline{\cup_{j=-\infty}^{\infty} V_j} = L^2[-\infty, \infty]$. (Thus, each $f \in L^2$ can be obtained a a limit of a Cauchy sequence $\{s_n : n = 1, 2, \cdots\}$ such that each $s_n \in V_{j_n}$ for some integer j_n.)*
3. *Separation: $\cap_{j=-\infty}^{\infty} V_j = \{0\}$, the subspace containing only the zero function. (Thus only the zero function is common to all subspaces V_j.)*
4. *Scale invariance: $f(t) \in V_j \Longleftrightarrow f(2t) \in V_{j+1}$.*
5. *Shift invariance of V_0: $f(t) \in V_0 \Longleftrightarrow f(t-k) \in V_0$ for all integers k.*
6. *ON basis: The set $\{\phi(t-k) : k = 0, \pm 1, \cdots\}$ is an ON basis for V_0.*

Here, the function $\phi(t)$ is called the scaling function (or the father wavelet).

REMARKS:

- The ON basis condition can be replaced by the (apparently weaker) condition that the translates of ϕ form a *Riesz basis*. This type of basis is most easily defined and understood from a frequency space viewpoint. We will show later that a ϕ determining a Riesz basis can be modified to a $\tilde{\phi}$ determining an ON basis.

- We can drop the ON basis condition and simply require that the integer translates of $\phi(t)$ form a basis for V_0. However, we will have to be precise about the meaning of this condition for an infinite dimensional space. We will take this up when we discuss frames. This will lead us to *biorthogonal wavelets*, in analogy with *biorthogonal* filter banks.

- The ON basis condition can be generalized in another way. It may be that there is no single function whose translates form an ON basis for V_0 but that there are m functions $\phi_1, \cdots, \phi_m$ with $m > 1$ such that the set $\{\phi_\ell(t-k) : \ell = 1, \cdots, m, \quad k = 0, \pm 1, \cdots\}$ is an ON basis for V_0. These generate *multiwavelets* and the associated filters are *multifilters*

- If the scaling function has finite support and satisfies the ON basis condition then it will correspond to a unitary FIR filter bank. If its support is not finite, however, it will still correspond to a unitary filter bank, but one that has Infinite Impulse Response (IIR). This means that the impulse response vectors $\mathbf{c}(n)$, $\mathbf{d}(n)$ have an infinite number of nonzero components.

EXAMPLES:

1. Piecewise constant functions. Here V_0 consists of the functions $f(t)$ that are constant on the unit intervals $k \le t < k+1$:

   $$f(t) = a_k \qquad \text{for } k \le t < k+1.$$

   This is exactly the Haar multiresolution analysis of the preceding section. The only change is that now we have introduced subspaces V_j for j negative. In this case the functions in V_{-n} for $n > 0$ are piecewise constant on the intervals $k \cdot 2^n \le t < (k+1) \cdot 2^n$. Note that if $f \in V_j$ for all integers j then f must be a constant. The only square integrable constant function is identically zero, so the separation requirement is satisfied.

2. Continuous piecewise linear functions. The functions $f(t) \in V_0$ are determined by their values $f(k)$ at the integer points, and are linear between each

pair of values:

$$f(t) = [f(k+1) - f(k)](t-k) + f(k) \quad \text{for } k \le t \le k+1.$$

Note that continuous piecewise linearity is invariant under integer shifts. Also if $f(t)$ is continuous piecewise linear on unit intervals, then $f(2t)$ is continuous piecewise linear on half-unit intervals. It isn't completely obvious, but a scaling function can be taken to be the *hat function*. The hat function $H(t)$ is the continuous piecewise linear function whose values on the integers are $H(k) = \delta_{0k}$, i.e., $H(0) = 1$ and $H(t)$ is zero on the other integers. The support of $H(t)$ is the open interval $-1 < t < 1$. Note that if $f \in V_0$ then we can write it uniquely in the form

$$f(t) = \sum_k f(k)H(t-k).$$

Although the sum could be infinite, at most 2 terms are nonzero for each t. Each term is linear, so the sum must be linear, and it agrees with $f(t)$ at integer times. All multiresolution analysis conditions are satisfied, except for the ON basis requirement. The integer translates of the hat function do define a basis for V_0 but it isn't ON because the inner product $(H(t), H(t-1)) \neq 0$. A scaling function does exist whose integer translates form an ON basis, but its support isn't compact.

3. Discontinuous piecewise linear functions. The functions $f(t) \in V_0$ are determined by their values and and left-hand limits $f(k), f(k-)$ at the integer points, and are linear between each pair of limit values:

$$f(t) = [f((k+1)-) - f(k)](t-k) + f(k) \quad \text{for } k \le t < k+1.$$

Each function $f(t)$ in V_0 is determined by the two values $f(k), f((k+1)-)$ in each unit subinterval $[k, k+1)$ and two scaling functions are needed:

$$\phi_1(t) = \begin{cases} 1 & \text{if } 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases} \qquad \phi_2(t) = \begin{cases} \sqrt{3}(1-2t) & \text{if } 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$f(t) = \sum_k \left(\frac{f(k)}{2}[\phi_1(t-k) + \frac{1}{\sqrt{3}}\phi_2(t-k)] \right.$$
$$\left. + \frac{f((k+1)-)}{2}[\phi_1(t-k)) - \frac{1}{\sqrt{3}}\phi_2(t-k)] \right).$$

The integer translates of $\phi_1(t), \phi_2(t)$ form an ON basis for V_0. These are multiwavelets and they correspond to multifilters.

4. Shannon resolution analysis. Here V_j is the space of band-limited signals $f(t)$ in $L^2[-\infty, \infty]$ with frequency band contained in the interval $[-2^j\pi, 2^j\pi]$. The nesting property is a consequence of the fact that if $f(t)$ has Fourier transform $\hat{f}(\lambda)$ then $f(2t)$ has Fourier transform $\frac{1}{2}\hat{f}(\frac{\lambda}{2})$. The function $\phi(t) = \mathrm{sinc}\ (t)$ is the scaling function. Indeed we have already shown that $||\phi|| = 1$ and The (unitary) Fourier transform of $\phi(t)$ is

$$\mathcal{F}\mathrm{sinc}\ (\lambda) = \begin{cases} \frac{1}{\sqrt{2\pi}} & \text{for } |\lambda| < \pi \\ 0 & \text{for } |\lambda| > \pi. \end{cases}$$

Thus the Fourier transform of $\phi(t-k)$ is equal to $e_k(\lambda) = \frac{e^{-ik\lambda}}{\sqrt{2\pi}}$ in the interior of the interval $[-\pi, \pi]$ and is zero outside this interval. It follows that the integer translates of $\mathrm{sinc}\ (t)$ form an ON basis for V_0. Note that the scaling function $\phi(t)$ does not have compact support in this case.

5. The Daubechies functions. We will see that each of the Daubechies unitary FIR filters D_n corresponds to a scaling function with compact support and an ON wavelets basis.

Just as in our study of the Haar multiresolution analysis, for a general multiresolution analysis we can define the functions

$$\phi_{jk}(t) = 2^{\frac{j}{2}}\phi(2^j t - k), \qquad k = 0, \pm 1, \pm 2, \cdots$$

and for fixed integer j they will form an ON basis for V_j. Since $V_0 \subset V_1$ it follows that $\phi \in V_1$ and ϕ can be expanded in terms of the ON basis $\{\phi_{1k}\}$ for V_1. Thus we have the *dilation equation*

$$\phi(t) = \sqrt{2}\sum_k \mathbf{c}(k)\phi(2t-k), \tag{7.22}$$

or, equivalently,

$$\phi(t) = 2\sum_{k=0}^{N} \mathbf{h}(k)\phi(2t-k). \tag{7.23}$$

Since the ϕ_{jk} form an ON set, the coefficient vector $\mathbf{c}$ must be a unit vector in ℓ^2,

$$\sum_k |\mathbf{c}(k)|^2 = 1. \tag{7.24}$$

We will soon show that $\phi(t)$ has support in the interval $[0, N]$ if and only if the only nonvanishing coefficients of $\mathbf{c}$ are $\mathbf{c}(0), \cdots, \mathbf{c}(N)$. Scaling functions with nonbounded support correspond to coefficient vectors with infinitely many nonzero terms. Since $\phi(t) \perp \phi(t-m)$ for all nonzero m, the vector $\mathbf{c}$ satisfies double-shift orthogonality:

$$(\phi_{00}, \phi_{0m}) = \sum_k \mathbf{c}(k)\overline{\mathbf{c}(k-2m)} = \delta_{0m}. \tag{7.25}$$

REMARK: For unitary FIR filters, double-shift orthogonality was associated with downsampling. For orthogonal wavelets it is associated with dilation.

From (7.23) we can easily prove

Lemma 39 *If the scaling function is normalized so that*

$$\int_{-\infty}^{\infty} \phi(t)dt = 1,$$

then $\sum_{k=0}^{N} \mathbf{c}(k) = \sqrt{2}$.

Also, note that the dilation equation (7.22) implies that

$$\phi_{jk}(t) = \sum_\ell \mathbf{c}(\ell - 2k)\phi_{j+1,\ell}(t), \tag{7.26}$$

which is the expansion of the V_j scaling basis in terms of the V_{j+1} scaling basis.

Just as in the special case of the Haar multiresolution analysis we can introduce the orthogonal complement W_j of V_j in V_{j+1}.

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

We start by trying to find an ON basis for the wavelet space W_0. Associated with the father wavelet $\phi(t)$ there must be a mother wavelet $w(t)$, with norm 1, and satisfying the *wavelet equation*

$$w(t) = \sqrt{2}\sum_k \mathbf{d}(k)\phi(2t-k), \tag{7.27}$$

or, equivalently,

$$w(t) = 2\sum_k \mathbf{h}_1(k)\phi(2t-k), \tag{7.28}$$

and such that w is orthogonal to all translations $\phi(t-k)$ of the father wavelet. We will further require that w is orthogonal to integer translations of itself. For the Haar scaling function $N = 1$ and $(\mathbf{h}_1(0), \mathbf{h}_1(1)) = (\frac{1}{2}, -\frac{1}{2})$. NOTE: In several of our examples we were able to identify the scaling subspaces, the scaling function and the mother wavelet explicitly. In general however, this won't be the case. Just as in our study of perfect reconstruction filter banks, we will determine conditions on the coefficient vectors $\mathbf{c}$ and $\mathbf{d}$ such that they could correspond to scaling functions and wavelets. We will solve these conditions and demonstrate that a solution defines a multiresolution analysis, a scaling function and a mother wavelet. Virtually the entire analysis will be carried out with the coefficient vectors; we shall seldom use the scaling and wavelet functions directly. Now back to our construction.

Since the ϕ_{jk} form an ON set, the coefficient vector $\mathbf{d}$ must be a unit vector in ℓ^2,

$$\sum_k |\mathbf{d}(k)|^2 = 1. \tag{7.29}$$

Moreover since $w(t) \perp \phi(t-m)$ for all m, the vector $\mathbf{d}$ satisfies double-shift orthogonality with $\mathbf{c}$:

$$(w, \phi_{0m}) = \sum_k \mathbf{c}(k)\overline{\mathbf{d}(k-2m)} = 0. \tag{7.30}$$

The requirement that $w(t) \perp w(t-m)$ for nonzero integer m leads to double-shift orthogonality of $\mathbf{d}$ to itself:

$$(w(t), w(t-m)) = \sum_k \mathbf{d}(k)\overline{\mathbf{d}(k-2m)} = \delta_{0m}. \tag{7.31}$$

From our earlier work on filters, we know that if the unit coefficient vector $\mathbf{c}$ is double-shift orthogonal then the coefficient vector $\mathbf{d}$ defined by taking the conjugate alternating flip automatically satisfies the conditions (7.30) and (7.31). Here,

$$\mathbf{d}(n) = (-1)^n\overline{\mathbf{c}(N-n)}. \tag{7.32}$$

This expression depends on N where the $\mathbf{c}$ vector for the low pass filter had $N+1$ nonzero components. However, due to the double-shift orthogonality obeyed by $\mathbf{c}$, the only thing about N that is necessary for $\mathbf{d}$ to exhibit double-shift orthogonality is that N be odd. Thus we will choose $N=-1$ and take $\mathbf{d}(n) = (-1)^n\overline{\mathbf{c}(-1-n)}$. (It will no longer be true that the support of $\mathbf{d}(n)$ lies in the set $n = 0, 1, \cdots N$ but for wavelets, as opposed to filters, this is not a problem.) Also, even though we originally derived this expression under the assumption that $\mathbf{c}$ and $\mathbf{d}$ had length N, it also works when $\mathbf{c} \in \ell^2$ has an infinite number of nonzero components. Let's check for example that $\mathbf{d}$ is orthogonal to $\mathbf{c}$:

$$S = \sum_k \mathbf{c}(k)\overline{\mathbf{d}(k-2m)} = \sum_k \mathbf{c}(k)(-1)^k \mathbf{c}(-1+2m-k)$$

Now set $\ell = -1 + 2m - k$ and sum over ℓ:

$$S = \sum_k \mathbf{c}(1-2m-\ell)(-1)^{\ell+1}\mathbf{c}(\ell) = -S.$$

Hence $S = 0$. Thus, once the scaling function is defined through the dilation equation, the wavelet $w(t)$ is determined by the wavelet equation (7.27) with $\mathbf{d}(n) = (-1)^n \overline{\mathbf{c}(-1-n)}$.

Once w has been determined we can define functions

$$w_{jk}(t) = 2^{\frac{j}{2}} w(2^j t - k)$$

$$j, k = 0, \pm 1, \pm 2, \cdots.$$

It is easy to prove the

Lemma 40

$$(w_{jk}, w_{j'k'}) = \delta_{jj'}\delta_{kk'}, \qquad (\phi_{jk}, w_{jk'}) = 0 \tag{7.33}$$

where $j, j', k, k' = 0, \pm 1, \cdots$.

The dilation and wavelet equations extend to:

$$\phi_{j\ell} = \sum_k \mathbf{c}(k - 2\ell)\phi_{j+1,k}(t), \tag{7.34}$$

$$w_{j\ell} = \sum_k \mathbf{d}(k - 2\ell)\phi_{j+1,k}(t), \tag{7.35}$$

Equations (7.34) and (7.35) fit exactly into our study of perfect reconstruction filter banks, particularly the the infinite matrix $\mathbf{H}_t = \begin{bmatrix} \mathbf{L} \\ \mathbf{B} \end{bmatrix}$ pictured in Figure 6.14. The rows of $\mathbf{H}_t$ were shown to be ON. Here we have replaced the finite impulse response vectors $\mathbf{c}(0), \cdots, \mathbf{c}(N)$ by possibly infinite vectors $\mathbf{c}(k)$ and have set $N = -1$ in the determination of $\mathbf{d}(k)$, but the proof of orthonormality of the rows of $\mathbf{H}_t$ still goes through, see Figure 7.8. Now, however, we have a different interpretation of the ON property. Note that the ℓth upper row vector is just the coefficient vector for the expansion of $\phi_{j\ell}(t)$ as a linear combination of

$$\mathbf{H}_t = \begin{bmatrix} \mathbf{L} \\ \mathbf{B} \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \mathbf{c}(0) & 0 & 0 & 0 & \cdot & \cdot \\ \cdot & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & \mathbf{c}(-1) & \cdot & \cdot \\ \cdot & \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \overline{\mathbf{c}}(-1) & \overline{\mathbf{c}}(0) & \overline{\mathbf{c}}(1) & \overline{\mathbf{c}}(2) & \cdot & \cdot \\ \cdot & \overline{\mathbf{c}}(-3) & -\overline{\mathbf{c}}(-2) & \overline{\mathbf{c}}(-1) & \overline{\mathbf{c}}(0) & \cdot & \cdot \\ \cdot & \overline{\mathbf{c}}(-5) & -\overline{\mathbf{c}}(-4) & \overline{\mathbf{c}}(-3) & -\overline{\mathbf{c}}(-2) & \overline{\mathbf{c}}(-1) & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot & \cdot \end{bmatrix}.$$

Figure 7.8: The wavelet $\mathbf{H}_t$ matrix

the ON basis vectors $\phi_{j+1,k}$. (Indeed the entry in upper row ℓ, column k is just the coefficient $\mathbf{c}(k-2\ell)$.) Similarly, the ℓth lower row vector is the coefficient vector for the expansion of $w_{j\ell}(t)$ as a linear combination of the basis vectors $\phi_{j+1,k}$ (and the entry in lower row ℓ, column k is the coefficient $\mathbf{d}(k-2\ell) = (-1)^k\overline{\mathbf{c}(2\ell-1-k)}$.)

In our study of perfect reconstruction filter banks we also showed that the columns of $\mathbf{H}_t$ were ON. This meant that the matrix was unitary and that its inverse was the transpose conjugate. We will check that the proof that the columns are ON goes through virtually unchanged, except that the sums may now be infinite and we have set $N=-1$. This means that we can solve equations (7.34) and (7.35) explicitly to express the basis vectors $\phi_{j+1,k}$ for V_{j+1} as linear combinations of the vectors $\phi_{j,k'}$ and $w_{jk''}$. The kth column of $\mathbf{H}_t$ is the coefficient vector for the expansion of $\phi_{j+1,k}$

Let's recall the conditions for orthonormality of the columns. The columns of $\mathbf{H}_t$ are of two types: even (containing only terms $\mathbf{c}(2n), \mathbf{d}(2n)$) and odd (containing only terms $\mathbf{c}(2n+1), \mathbf{d}(2n+1)$). Thus the requirement that the column vectors of $\mathbf{H}_t$ are ON reduces to 3 types of identities:

$$\begin{aligned} \text{even} - \text{even}: \quad & \sum_\ell \mathbf{c}(2\ell)\overline{\mathbf{c}(2k+2\ell)} \\ + \quad & \sum_\ell \mathbf{d}(2\ell)\overline{\mathbf{d}(2k+2\ell)} = \delta_{k0} \end{aligned} \qquad (7.36)$$

$$\begin{aligned}\text{odd}-\text{odd}: \quad & \sum_\ell \mathbf{c}(2\ell+1)\overline{\mathbf{c}}(2k+2\ell+1) \\ + \quad & \sum_\ell \mathbf{d}(2\ell+1)\overline{\mathbf{d}(2k+2\ell+1)} = \delta_{k0} \qquad (7.37) \\ \text{odd}-\text{even}: \quad & \sum_\ell \mathbf{c}(2\ell+1)\overline{\mathbf{c}(2k+2\ell)} \\ + \quad & \sum_\ell \mathbf{d}(2\ell+1)\overline{\mathbf{d}(2k+2\ell)} = 0. \qquad (7.38)\end{aligned}$$

Theorem 43 *If* $\mathbf{c}$ *satisfies the double shift orthogonality condition and the filter* $\mathbf{d}$ *is determined by the conjugate alternating flip*

$$\mathbf{d}(n) = (-1)^n\overline{\mathbf{c}(-1-n)},$$

then the columns of $\mathbf{H}_t$ *are orthonormal.*

PROOF: The proof is virtually identical to that of the corresponding theorem for filter banks. We just set $N=-1$ everywhere. For example the even-even case computation is

$$\sum_\ell \mathbf{d}(2\ell)\overline{\mathbf{d}(2k+2\ell)} = \sum_\ell \overline{\mathbf{c}(-1-2\ell)}\mathbf{c}(-1-2k-2\ell)$$

$$= \sum_s \mathbf{c}(2s+1)\overline{\mathbf{c}(2s+2k+1)}.$$

Thus

$$\sum_\ell \mathbf{c}(2\ell)\overline{\mathbf{c}(2k+2\ell)} + \sum_\ell \mathbf{d}(2\ell)\overline{\mathbf{d}(2k+2\ell)}$$

$$= \sum_n \mathbf{c}(n)\overline{\mathbf{c}(n+2k)} = \delta_{k0}.$$

Q.E.D.

Now we define functions $\phi'_{j+1,k}(t)$ in V_{j+1} by

$$\phi'_{j+1,s} = \sum_h \left(\overline{\mathbf{c}(s-2h)}\phi_{jh} + \overline{\mathbf{d}(s-2h)}w_{jh}\right),$$

Substituting the expansions

$$\phi_{jh} = \sum_k \mathbf{c}(k-2h)\phi_{j+1,k},$$

$$w_{jh} = \sum_k \mathbf{d}(k-2h)\phi_{j+1,k},$$

into the right-hand side of the first equation we find

$$\phi'_{j+1,s} = \sum_{hk} \left(\overline{\mathbf{c}(s-2h)}\mathbf{c}(k-2h) + \overline{\mathbf{d}(s-2h)}\mathbf{d}(k-2h) \right) \phi_{j+1,k} = \phi_{j+1,s},$$

as follows from the even-even, odd-even and odd-odd identities above. Thus

$$\phi_{j+1,s} = \sum_h \left(\overline{\mathbf{c}(s-2h)}\phi_{jh} + \overline{\mathbf{d}(s-2h)}w_{jh} \right), \qquad (7.39)$$

and we have inverted the expansions

$$\phi_{j\ell} = \sum_k \mathbf{c}(k-2\ell)\phi_{j+1,k}(t), \qquad (7.40)$$

$$w_{j\ell} = \sum_k \mathbf{d}(k-2\ell)\phi_{j+1,k}(t), \qquad (7.41)$$

Thus the set $\{\phi_{jk}, w_{jk'}\}$ is an alternate ON basis for V_{j+1} and we have the

Lemma 41 *The wavelets* $\{w_{jk} : k = 0, \pm 1, \cdots\}$ *form an ON basis for* W_j.

To get the wavelet expansions for functions $f \in L^2$ we can now follow the steps in the construction for the Haar wavelets. The proofs are virtually identical. Since $V_j \oplus W_j = V_{j+1}$ for all $j \geq 0$, we can iterate on j to get $V_{j+1} = W_j \oplus V_j = W_j \oplus W_{j-1} \oplus V_{j-1}$ and so on. Thus

$$V_{j+1} = W_j \oplus W_{j-1} \oplus \cdots \oplus W_1 \oplus W_0 \oplus V_0.$$

and any $s \in V_{j+1}$ can be written uniquely in the form

$$s = \sum_{k=0}^{j} w_k + s_0 \qquad \text{where } w_k \in W_k,\ s_0 \in V_0.$$

Theorem 44

$$L^2[-\infty, \infty] = V_j \oplus \sum_{k=j}^{\infty} W_k = V_j \oplus W_j \oplus W_{j+1} \oplus \cdots,$$

so that each $f(t) \in L^2[-\infty, \infty]$ *can be written uniquely in the form*

$$f = f_j + \sum_{k=j}^{\infty} w_k, \qquad w_k \in W_k,\ f_j \in V_j. \qquad (7.42)$$

We have a family of new ON bases for $L^2[-\infty,\infty]$, one for each integer j:

$$\{\phi_{jk}, w_{j'k'}: \qquad j'=j,j+1,\cdots, \quad \pm k,\pm k'=0,1,\cdots\}.$$

Let's consider the space V_j for fixed j. On one hand we have the scaling function basis

$$\{\phi_{j,k}: \qquad \pm k=0,1,\cdots\}.$$

Then we can expand any $f_j\in V_j$ as

$$f_j=\sum_{k=-\infty}^{\infty} a_{j,k}\phi_{j,k}. \tag{7.43}$$

On the other hand we have the wavelets basis

$$\{\phi_{j-1,k}, w_{j-1,k'}: \qquad \pm k,\pm k'=0,1,\cdots\}$$

associated with the direct sum decomposition

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

Using this basis we can expand any $f_j\in V_j$ as

$$f_j=\sum_{k'=-\infty}^{\infty} b_{j-1,k'}w_{j-1,k'}+\sum_{k=-\infty}^{\infty} a_{j-1,k}\phi_{j-1,k}. \tag{7.44}$$

If we substitute the relations

$$\phi_{j-1,\ell}=\sum_k \mathbf{c}(k-2\ell)\phi_{jk}(t), \tag{7.45}$$

$$w_{j-1,\ell}=\sum_k \mathbf{d}(k-2\ell)\phi_{j,k}(t), \tag{7.46}$$

into the expansion (7.43) and compare coefficients of $\phi_{j,\ell}$ with the expansion (7.44), we obtain the fundamental recursions

$$\begin{aligned} \text{Averages(lowpass)} \quad & a_{j-1,k}=\textstyle\sum_n \mathbf{c}(n-2k)a_{jn} && (7.47)\\ \text{Differences(highpass)} \quad & b_{j-1,k}=\textstyle\sum_n \mathbf{d}(n-2k)a_{jn}. && (7.48)\end{aligned}$$

These equations link the wavelets with the unitary filter bank. Let $\mathbf{x}(k)=a_{jk}$ be a discrete signal. The result of passing this signal through the (normalized and time reversed) filter $\mathbf{C}^T$ and then downsampling is $\mathbf{y}(k)=(\downarrow 2)\mathbf{C}^T*\mathbf{x}(k)=a_{j-1,k}$,

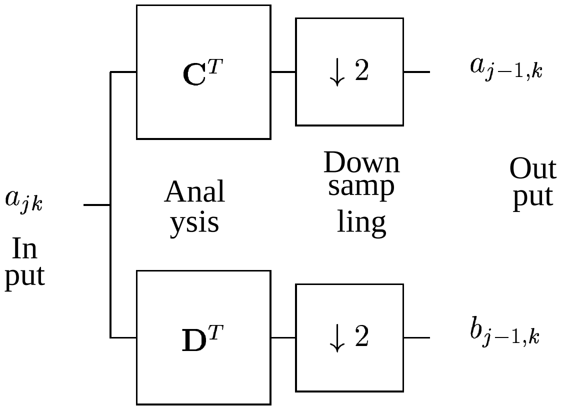

Figure 7.9: Wavelet Recursion

where $a_{j-1,k}$ is given by (7.47). Similarly, the result of passing the signal through the (normalized and time-reversed) filter $\mathbf{D}^T$ and then downsampling is $\mathbf{z}(k) = (\downarrow 2)\mathbf{D}^T * \mathbf{x}(k) = b_{j-1,k}$, where $b_{j-1,k}$ is given by (7.48).

The picture, in complete analogy with that for Haar wavelets, is in Figure 7.9.

We can iterate this process by inputting the output $a_{j-1,k}$ of the high pass filter to the filter bank again to compute $a_{j-2,k}, b_{j-2,k}$, etc. At each stage we save the wavelet coefficients $b_{j'k'}$ and input the scaling coefficients $a_{j'k'}$ for further processing, see Figure 7.10. The output of the final stage is the set of scaling coefficients a_{0k}, assuming that we stop at $j = 0$. Thus our final output is the complete set of coefficients for the wavelet expansion

$$f_j = \sum_{j'=0}^{j-1} \sum_{k=-\infty}^{\infty} b_{j'k} w_{j'k} + \sum_{k=-\infty}^{\infty} a_{0k}\phi_{0k},$$

based on the decomposition

$$V_j = W_{j-1} \oplus W_{j-2} \oplus \cdots \oplus W_1 \oplus W_0 \oplus V_0.$$

To derive the synthesis filter bank recursion we can substitute the inverse relation

$$\phi_{j,s} = \sum_h \left(\overline{\mathbf{c}}(s-2h)\phi_{j-1,h} + \overline{\mathbf{d}}(s-2h)w_{j-1,h} \right), \tag{7.49}$$

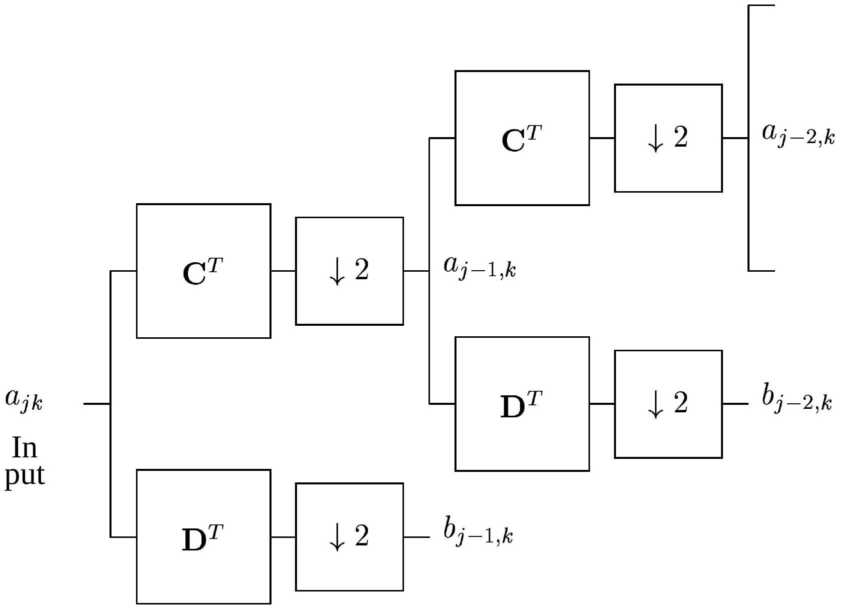

Figure 7.10: General Fast Wavelet Transform

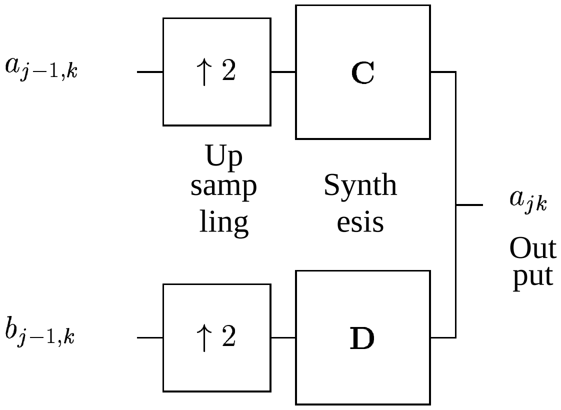

Figure 7.11: Wavelet inversion

into the expansion (7.44) and compare coefficients of $\phi_{j-1,\ell}, w_{j-1,\ell}$ with the expansion (7.43) to obtain the inverse recursion

$$a_{j,n} = \sum_k \mathbf{c}(2k-n)a_{j-1,k} + \sum_k \mathbf{d}(2k-n)b_{j-1,k}. \tag{7.50}$$

This is exactly the output of the synthesis filter bank shown in Figure 7.11.

Thus, for level j the full analysis and reconstruction picture is Figure 7.12.

In analogy with the Haar wavelets discussion, for any $f(t) \in L^2[-\infty, \infty]$ the scaling and wavelets coefficients of f are defined by

$$\begin{aligned} a_{jk} &= (f, \phi_{jk}) = 2^{j/2} \int_{-\infty}^{\infty} f(t)\overline{\phi(2^j t - k)}dt, \\ b_{jk} &= (f, w_{jk}) = 2^{j/2} \int_{-\infty}^{\infty} f(t)\overline{w(2^j t - k)}dt \end{aligned} \tag{7.51}$$

7.2.1 Wavelet Packets

The wavelet transform of the last section has been based on the decomposition $V_j = W_{j-1} \oplus V_{j-1}$ and its iteration. Using the symbols a_j, b_j (with the index k suppressed) for the projection of a signal f on the subspaces V_j, W_j, respectively, we have the tree structure of Figure 7.13 where we have gone down three levels in the recursion. However, a finer resolution is possible. We could also use our

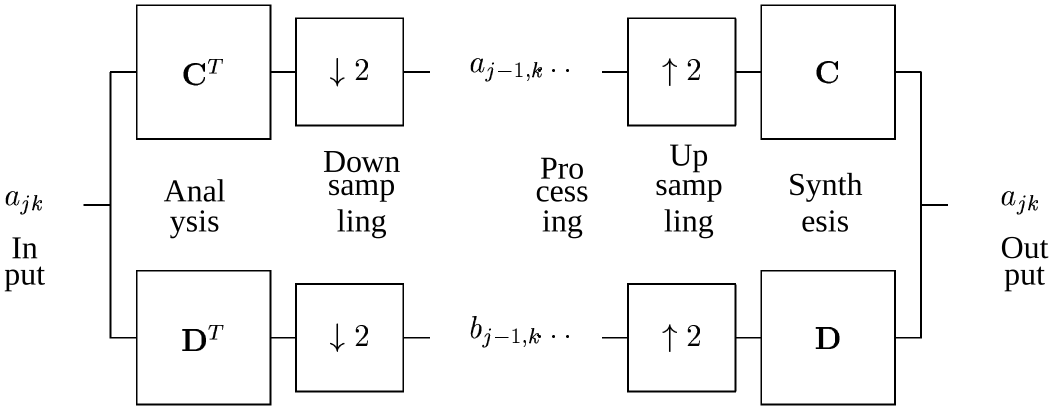

Figure 7.12: General Fast Wavelet Transform and Inversion

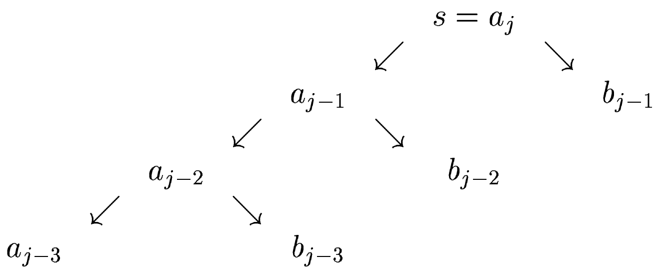

Figure 7.13: General Fast Wavelet Transform Tree

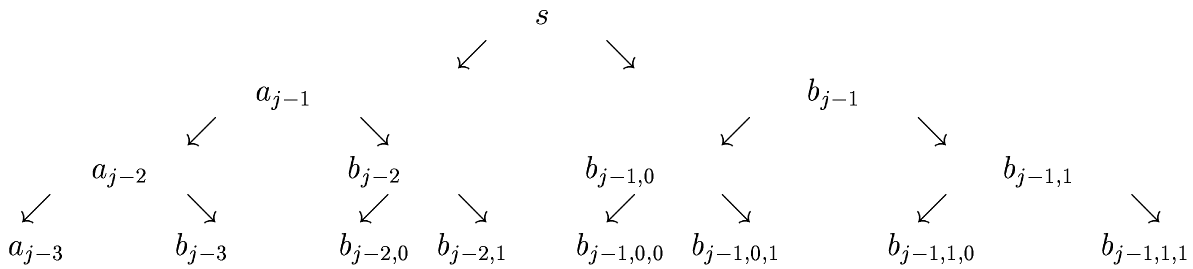

Figure 7.14: Wavelet Packet Tree

low pass and high pass filters to decompose the wavelet spaces W_j into a direct sum of low frequency and high frequency subspaces: $W_j = W_{j,0} \oplus W_{j,1}$. The new ON basis for this decomposition could be obtained from the wavelet basis $w_{jk}(t)$ for W_j exactly as the basis for the decomposition $V_j = W_{j-1} \oplus V_{j-1}$ was obtained from the scaling basis $\phi_{jk}(t)$ for V_j: $w_{j\ell,1}(t) = \sum_{k=0}^{N} \mathbf{d}(k-2\ell)w_{jk}(t)$ and $w_{j\ell,0}(t) = \sum_{k=0}^{N} \mathbf{c}(k-2\ell)w_{jk}(t)$. Similarly, the new high and low frequency wavelet subspaces so obtained could themselves be decomposed into a direct sum of high and low pass subspaces, and so on. The wavelet transform algorithm (and its inversion) would be exactly as before. The only difference is that the algorithm would be applied to the b_{jk} coefficients, as well as the a_{jk} coefficients. Now the picture (down three levels) is the complete (wavelet packet) Figure 7.14. With wavelet packets we have a much finer resolution of the signal and a greater variety of options for decomposing it. For example, we could decompose s as the sum of the 8 terms at level three: $s = a_{j-3} + b_{j-3} + b_{j-2,0} + b_{j-2,1} + b_{j-1,0,0} + b_{j-1,0,1} + b_{j-1,1,0} + b_{j-1,1,1}$. A hybrid would be $s = a_{j-1} + b_{j-1,0} + b_{j-1,1,0} + b_{j-1,1,1}$. The tree structure for this algorithm is the same as for the FFT. The total number of multiplications involved in analyzing a signal at level J all the way down to level 0 is of the order $J2^J$, just as for the Fast Fourier Transform.

7.3 Sufficient conditions for multiresolution analysis

We are in the process of constructing a family of continuous scaling functions with compact support and such that the integer translates of each scaling function form an ON set. Even if this construction is successful, it isn't yet clear that each of these scaling functions will in fact generate a basis for $L^2[-\infty,\infty]$, i.e. that any function in L^2 can be approximated by these wavelets. The following results will show that indeed such scaling functions determine a multiresolution analysis.

First we collect our assumptions concerning the scaling function $\phi(t)$.

- Condition A: Suppose that $\phi(t)$ is continuous with compact support on the real line and that it satisfies the orthogonality conditions $(\phi_{0k},\phi_{0\ell}) = \int \phi(t-k)\overline{\phi(t-\ell)}dt = \delta_{k\ell}$ in V_0. Let V_j be the subspace of $L^2[-\infty,\infty]$ with ON basis $\{\phi_{jk} : k = 0, \pm 1, \cdots\}$ where $\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k)$.

- Condition B: Suppose ϕ satisfies the normalization condition $\int \phi(t)dt = 1$ and the dilation equation

$$\phi = \sqrt{2}\sum_{k=0}^{N} \mathbf{c}(k)\phi(2t-k)$$

for finite N.

Lemma 42 *If Condition A is satisfied then for all $f \in V_0$ and for all t, there is a constant K such that*

$$|f(t)| \le K||f||.$$

PROOF: If $f \in V_0$ then we have $f(t) = \sum_k c_k\phi_{0k}(t)$. We can assume pointwise equality as well as Hilbert space equality, because for each t only a finite number of the continuous functions $\phi(t-k)$ are nonzero. Since $c_k = (f, \phi_{0k})$ we have

$$f(t) = \int h(t,s)f(s)ds, \quad \text{where } h(t,s) = \sum_k \phi(t-k)\overline{\phi(s-k)}.$$

Again, for any t, s only a finite number of terms in the sum for $h(t,s)$ are nonzero. For fixed t the kernel $h(t,s)$ belongs to the inner product space of square integrable functions in s. The norm square of h in this space is

$$||h(t,\cdot)||_s^2 = \sum_k |\phi(t-k)|^2 \le K^2$$

for some positive constant K. This is because only a finite number of the terms in the k-sum are nonzero and $\phi(t)$ is a bounded function. Thus by the Schwarz inequality we have

$$|f(t)| = |(k(t,\cdot), f)_s| \le ||h(t,\cdot)||||f|| \le K||f||.$$

Q.E.D.

Theorem 45 *If Condition A is satisfied then the separation property for multiresolution analysis holds:* $\cap_{j=-\infty}^{\infty} V_j = \{0\}$.

PROOF: Suppose $f \in V_{-j}$. This means that $f(2^j t) \in V_0$. By the lemma, we have

$$|f(2^j t)| \leq K||f(2^j \cdot)|| = K2^{-j/2}||f||.$$

If $f \in V_{-j}$ for all j then $f(t) \equiv 0$. Q.E.D.

Theorem 46 *If both Condition A and Condition B are satisfied then the density property for multiresolution analysis holds:* $\overline{\cup_{j=-\infty}^{\infty} V_j} = L^2[-\infty, \infty]$.

PROOF: Let $R_{ab}(t)$ be a rectangular function:

$$R_{ab}(t) = \begin{cases} 1 & \text{for } a \leq t \leq b \\ 0 & \text{otherwise.} \end{cases}$$

for $a < b$. We will show that $R_{ab} \in \overline{\cup_{j=-\infty}^{\infty} V_j}$. Since every step function is a linear combination of rectangular functions, and since the step functions are dense in $L^2[-\infty, \infty]$, this will prove the theorem. Let $P_j R_{ab}$ be the orthogonal projection of R_{ab} on the space V_j. Since $\{\phi_{jk}\}$ is an ON basis for V_j we have

$$P_j R_{ab} = \sum_k c_k \phi_{jk}.$$

We want to show that $||R_{ab} - P_j R_{ab}|| \to 0$ as $j \to +\infty$. Since $R_{ab} - P_j R_{ab} \perp V_j$ we have

$$||R_{ab}||^2 = ||R_{ab} - P_j R_{ab}||^2 + ||P_j R_{ab}||^2,$$

so it is sufficient to show that $||P_j R_{ab}||^2 \to ||R_{ab}||^2$ as $j \to +\infty$. Now

$$||P_j R_{ab}||^2 = \sum_k |c_k|^2 = \sum_k |(R_{ab}, \phi_{jk})|^2 = 2^j \sum |\int_a^b \phi(2^j t - k) dt|^2$$

so

$$||P_j R_{ab}||^2 = 2^{-j} \sum_k |\int_{2^j a}^{2^j b} \phi(t - k) dt|^2.$$

Now the support of $\phi(t)$ is contained in some finite interval with integer endpoints $m_1 < m_2$: $m_1 \leq t \leq m_2$. For each integral in the summand there are three possibilities:

1. The intervals $[2^j a, 2^j b]$ and $[m_1, m_2]$ are disjoint. In this case the integral is 0.

2. $[m_1, m_2] \subset [2^j a, 2^j b]$. In this case the integral is 1.

3. The intervals $[2^j a, 2^j b]$ and $[m_1, m_2]$ partially overlap. As j gets larger and larger this case is more and more infrequent. Indeed if $ab \neq 0$ this case won't occur at all for sufficiently large j. It can only occur if say, $a = 0, m_1 \leq 0, m_2 > 0$. For large j the number of such terms would be fixed at $|m_1|$. In view of the fact that each integral squared is multiplied by 2^{-j} the contribution of these boundary terms goes to 0 as $j \to +\infty$.

Let $N_j(a,b)$ equal the number of integers between $2^j a$ and $2^j b$. Clearly $N_j(a,b) \sim 2^j(b-a)$ and $2^{-j}N_j(a,b) \to b-a$ as $j \to +\infty$. Hence

$$\lim_{j \to +\infty} ||P_j R_{a,b}||^2 = b - a = \int_a^b 1 dt = ||R_{ab}||^2.$$

Q.E.D.

7.4 Lowpass iteration and the cascade algorithm

We have come a long way in our study of wavelets, but we still have no concrete examples of father wavelets other than a few that have been known for almost a century. We now turn to the problem of determining new multiresolution structures. Up to now we have been accumulating *necessary conditions* that must be satisfied for a multiresolution structure to exist. Our focus has been on the coefficient vectors **c** and **d** of the dilation and wavelet equations. Now we will gradually change our point of view and search for more restrictive *sufficient conditions* that will guarantee the existence of a multiresolution structure. Further we will study the problem of actually computing the scaling function and wavelets. In this section we will focus on the time domain. In the next section we will go to the frequency domain, where new insights emerge. Our work with Daubechies filter banks will prove invaluable since they are all associated with wavelets.

Our main focus will be on the dilation equation

$$\phi(t) = \sqrt{2}\sum_k \mathbf{c}(k)\phi(2t-k). \tag{7.52}$$

We have already seen that if we have a scaling function satisfying this equation, then we can define **d** from **c** by a conjugate alternating flip and use the wavelet equation to generate the wavelet basis. Our primary interest is in scaling functions ϕ with support in a finite interval.

If ϕ has finite support then by translation in time if necessary, we can assume that the support is contained in the interval $[0, N)$. With such a $\phi(t)$ note that even though the right-hand side of (7.52) could conceivably have an infinite number of nonzero $\mathbf{c}(k)$, for fixed t there are only a finite number of nonzero terms. Suppose the support of $\mathbf{c}$ is contained in the interval $[N_1, N_2]$ (which might be infinite). Then the support of the right-hand side is contained in $[\frac{N_1}{2}, \frac{N+N_2}{2})$. Since the support of both sides is the same, we must have $N_1 = 0, N_2 = N$. Thus $\mathbf{c}$ has only $N+1$ nonzero terms $\mathbf{c}(0), \mathbf{c}(1), \cdots, \mathbf{c}(N)$. Further, N must be odd, in order that $\mathbf{c}$ satisfy the double-shift orthogonality conditions.

Lemma 43 *If the scaling function $\phi(t)$ (corresponding to a multiresolution analysis) has compact support on $[0, N)$, then $\mathbf{c}(k)$ also has compact support over $0 \leq k \leq N$.*

Recall that $\mathbf{c}$ also must obey the double-shift orthogonality conditions

$$\sum_k \mathbf{c}(k)\overline{\mathbf{c}}(k-2m) = \delta_{0m}$$

and the compatibility condition

$$\sum_{k=0}^{N} \mathbf{c}(k) = \sqrt{2}$$

between the unit area normalization $\int \phi(t)dt = 1$ of the scaling function and the dilation equation.

One way to try to determine a scaling function $\phi(t)$ from the impulse response vector $\mathbf{c}$ is to iterate the lowpass filter $\mathbf{C}$. That is, we start with an initial guess $\phi^{(0)}(t)$, the box function on $[0, 1)$, and then iterate

$$\phi^{(i+1)}(t) = \sqrt{2}\sum_{k=0}^{N} \mathbf{c}(k)\phi^{(i)}(2t-k) \tag{7.53}$$

for $i = 1, 2, \cdots$. Note that $\phi^{(i)}(t)$ will be a piecewise constant function, constant on intervals of length $\frac{1}{2^j}$. If $\lim_{i\to\infty} \phi^{(i)}(t) \equiv \phi(t)$ exists for each t then the limit function satisfies the dilation equation (7.52). This is called the *cascade algorithm*, due to the iteration by the low pass filter.

Of course we don't know in general that the algorithm will converge. (We will find a sufficient condition for convergence when we look at this algorithm in the frequency domain.) For the moment, let's look at the implications of uniform convergence on $[-\infty, \infty]$ of the sequence $\phi^{(i)}(t)$ to $\phi(t)$.

First of all, the support of ϕ is contained in the interval $[0, N)$. To see this note that, first, the initial function $\phi^{(0)}$ has support in $[0, 1)$. After filtering once, we see that the new function $\phi^{(1)}$ has support in $[0, \frac{1+N}{2})$. Iterating, we see that $\phi^{(i)}$ has support in $[0, \frac{1+[2^i-1]N}{2^i})$.

Note that at level $i = 0$ the scaling function and associated wavelets are orthonormal:

$$(w^{(0)}_{jk}, w^{(0)}_{j'k'}) = \delta jj'\delta_{kk'}, \qquad (\phi^{(0)}_{jk}, w^{(0)}_{jk'}) = 0$$

where $j, j', \pm k, \pm k' = 0, 1, \cdots$. (Of course it is *not* true in general that $\phi^{(0)}_{jk} \perp \phi^{(0)}_{j'k'}$ for $j \neq j'$.) These are just the orthogonality relations for the Haar wavelets. This orthogonality is maintained through each iteration, and if the cascade algorithm converges uniformly, it applies to the limit function ϕ:

Theorem 47 *If the cascade algorithm converges uniformly in t then the limit function $\phi(t)$ and associated wavelet $w(t)$ satisfy the orthogonality relations*

$$(w_{jk}, w_{j'k'}) = \delta_{jj'}\delta_{kk'}, \qquad (\phi_{jk}, w_{jk'}) = 0, \qquad (\phi_{jk}, \phi_{jk'}) = \delta_{kk'}$$

where $j, j', \pm k, \pm k' = 0, 1, \cdots$.

PROOF: There are only three sets of identities to prove:

$$1. \qquad \int_{-\infty}^{\infty} \phi(t-n)\overline{\phi(t-m)}dt = \delta_{nm}$$

$$2. \qquad \int_{-\infty}^{\infty} \phi(t-n)\overline{w(t-m)}dt = 0$$

$$3. \qquad \int_{-\infty}^{\infty} w(t-n)\overline{w(t-m)}dt = \delta_{nm}.$$

The rest are immediate.

1. We will use induction. If 1. is true for the function $\phi^{(i)}(t)$ we will show that it is true for the function $\phi^{(i+1)}(t)$. Clearly it is true for $\phi^{(0)}(t)$. Now

$$\int_{-\infty}^{\infty} \phi^{(i+1)}(t-n)\overline{\phi^{(i+1)}(t-m)}dt = (\phi^{(i+1)}_{0n}, \phi^{(i+1)}_{0m})$$

$$= (\sum_k \mathbf{c}(k)\phi^{(i)}_{1,2n+k}, \sum_\ell \mathbf{c}(\ell)\phi^{(i)}_{1,2m+\ell})$$

$$= \sum_{k\ell} \mathbf{c}(k)\overline{\mathbf{c}(\ell)}(\phi^{(i)}_{1,2n+k}, \phi^{(i)}_{1,2m+\ell}) = \sum_k \mathbf{c}(k)\overline{\mathbf{c}(k-2(m-n))} = \delta_{nm}.$$

Since the convergence is uniform and $\phi(t)$ has compact support, these orthogonality relations are also valid for $\phi(t)$.

2.

$$\int_{-\infty}^{\infty} \phi^{(i+1)}(t-n)\overline{w^{(i+1)}(t-m)}dt = (\phi^{(i+1)}_{0n}, w^{(i+1)}_{0m})$$

$$= (\sum_k \mathbf{c}(k)\phi^{(i)}_{1,2n+k}, \sum_\ell \mathbf{d}(\ell)w^{(i)}_{1,2m+\ell}$$

$$= \sum_{k\ell} \mathbf{c}(k)\overline{\mathbf{d}(\ell)}(\phi^{(i)}_{1,2n+k}, w^{(i)}_{1,2m+\ell}) = \sum_k \mathbf{c}(k)\overline{\mathbf{d}(k-2(m-n))} = 0,$$

because of the double-shift orthogonality of $\mathbf{c}$ and $\mathbf{d}$.

3.

$$\int_{-\infty}^{\infty} w^{(i+1)}(t-n)\overline{w^{(i+1)}(t-m)}dt = (w^{(i+1)}_{0n}, w^{(i+1)}_{0m})$$

$$= \sum_k \mathbf{d}(k)\overline{\mathbf{d}(k-2(m-n))} = \delta_{nm},$$

because of the double-shift orthonormality of $\mathbf{d}$.

Q.E.D.

Note that most of the proof of the theorem doesn't depend on convergence. It simply relates properties at the ith recursion of the cascade algorithm to the same properties at the $(i+1)$-st recursion.

Corollary 13 *If the the orthogonality relations*

$$(w^{(i)}_{jk}, w^{(i)}_{j'k'}) = \delta_{jj'}\delta_{kk'}, \qquad (\phi^{(i)}_{jk}, w^{(i)}_{jk'}) = 0, \qquad (\phi^{(i)}_{jk}, \phi^{(i)}_{jk'}) = \delta_{kk'}$$

where $j, j', \pm k, \pm k' = 0, 1, \cdots$ are valid at the ith recursion of the cascade algorithm they are also valid at the $(i+1)$-st recursion.

7.5 Scaling Function by recursion. Evaluation at dyadic points

We continue our study of topics related to the cascade algorithm. We are trying to characterize multiresolution systems with scaling functions $\phi(t)$ that have support in the interval $[0, N)$ where N is an odd integer. The low pass filter that characterizes the system is $\mathbf{c}(0), \cdots, \mathbf{c}(N)$. One of the beautiful features of the dilation equation is that it enables us to compute explicitly the values $\phi(\frac{k}{2^j})$ for all j, k, i.e., at all dyadic points. Each value can be obtained as a result of a finite (and easily determined) number of passes through the low pass filter. The dyadic points are dense in the reals, so if we know that ϕ exists and is continuous, we will have determined it completely.

The hardest step in this process is the first one. The dilation equation is

$$\phi(t) = \sqrt{2}\sum_{k=0}^{N} \mathbf{c}(k)\phi(2t-k). \tag{7.54}$$

If $\phi(t)$ exists, it is zero outside the interval $0 \le t < N$, so we can restrict our attention to the values of $\phi(t)$ on $[0, N)$. We first try to compute $\phi(t)$ on the integers $t = 0, 1, \cdots, N-1$. Substituting these values one at a time into (7.54) we obtain the system of equations

$$\begin{bmatrix} \phi(0) \\ \phi(1) \\ \phi(2) \\ \phi(3) \\ \phi(4) \\ \cdots \\ \phi(N-2) \\ \phi(N-1) \end{bmatrix} = \sqrt{2} \begin{bmatrix} \mathbf{c}(0) & 0 & & & & & & \\ \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & 0 & & \cdots & & \\ \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & \cdots & & \\ \mathbf{c}(6) & \mathbf{c}(5) & \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \cdots & & \\ \mathbf{c}(8) & \mathbf{c}(7) & \mathbf{c}(6) & \mathbf{c}(5) & \mathbf{c}(4) & \cdots & & \\ & & & \cdots & & \cdots & & \\ & & & \cdots & & \cdots & \mathbf{c}(N-2) & \mathbf{c}(N-3) \\ & & & & & \cdots & \mathbf{c}(N) & \mathbf{c}(N-1) \end{bmatrix} \begin{bmatrix} \phi(0) \\ \phi(1) \\ \phi(2) \\ \phi(3) \\ \phi(4) \\ \cdots \\ \phi(N-2) \\ \phi(N-1) \end{bmatrix},$$

or

$$\Phi(0) = \mathbf{m}(0)\Phi(0). \tag{7.55}$$

This says that $\Phi(0)$ is an eigenvector of the $N \times N$ matrix $\mathbf{m}(0)$, with eigenvalue 1. If 1 is in fact an eigenvalue of $\mathbf{m}(0)$ then the homogeneous system of equations (7.55) can be solved for $\Phi(0)$ by Gaussian elimination.

We can show that $\mathbf{m}(0)$ always has 1 as an eigenvalue, so that (7.55) always has a nonzero solution. We need to recall from linear algebra that λ is an eigenvalue of the $N \times N$ matrix $\mathbf{m}(0)$ if and only if it is a solution of the characteristic equation

$$\det[\mathbf{m}(0) - \lambda I] = 0. \tag{7.56}$$

Since the determinant of a square matrix equals the determinant of its transpose, we have that (7.56) is true if and only if

$$\det[\mathbf{m}^{\mathrm{tr}}(0) - \lambda I] = 0.$$

Thus $\mathbf{m}(0)$ has 1 as an eigenvalue if and only if $\mathbf{m}^{\mathrm{tr}}(0)$ has 1 as an eigenvalue. I claim that the column vector $(1, 1, \cdots, 1)$ is an eigenvector of $\mathbf{m}^{\mathrm{tr}}(0)$. Note that the column sum of each of the 1st, 3rd 5th, ... columns of $\mathbf{m}(0)$ is $\sqrt{2}\sum_k \mathbf{c}(2k)$, whereas the column sum of each of the even-numbered columns is $\sqrt{2}\sum_k \mathbf{c}(2k+1)$. However, it is a consequence of the double-shift orthogonality conditions

$$\sum_k \mathbf{c}(k)\overline{\mathbf{c}}(k-2m) = \delta_{0m}$$

and the compatibility condition

$$\sum_{k=0}^{N} \mathbf{c}(k) = \sqrt{2}$$

that each of those sums is equal to 1. Thus the column sum of each of the columns of $\mathbf{m}(0)$ is 1, which means that the row sum of each of the rows of $\mathbf{m}^{\mathrm{tr}}(0)$ is 1, which says precisely that the column vector $(1, 1, \cdots, 1)$ is an eigenvector of $\mathbf{m}^{\mathrm{tr}}(0)$ with eigenvalue 1.

The identities

$$\sqrt{2}\sum_k \mathbf{c}(2k) = \sqrt{2}\sum_k \mathbf{c}(2k+1) = 1$$

can be proven directly from the above conditions (and I will assign this as a homework problem). An indirect but simple proof comes from these equations in frequency space. Then we have the Fourier transform

$$C(\omega) = \sum_k \mathbf{c}(k)e^{-ik\omega}.$$

The double-shift orthogonality condition is now expressed as

$$|C(\omega)|^2 + |C(\omega+\pi)|^2 = 2. \tag{7.57}$$

The compatibility condition says

$$C(0) = \sum_k \mathbf{c}(k) = \sqrt{2}.$$

It follows from (7.57) that

$$C(\pi) = 0 = \sum_k (-1)^k \mathbf{c}(k).$$

from $C(\pi) = 0$ we have $\sum_k \mathbf{c}(2k) = \sum_k \mathbf{c}(2k+1)$. So the column sums are the same. Then from $C(0) = \sqrt{2}$ we get our desired result.

Now that we can compute the scaling function on the integers (*up to a constant multiple; we shall show how to fix the normalization constant shortly*) we can proceed to calculate $\phi(t)$ at all dyadic points $t = \frac{k}{2^j}$. The next step is to compute $\phi(t)$ on the half-integers $t = 1/2, 3/2, \cdots, N-1/2$. Substituting these values one at a time into (7.54) we obtain the system of equations

$$\begin{bmatrix} \phi(\frac{1}{2}) \\ \phi(\frac{3}{2}) \\ \phi(\frac{5}{2}) \\ \phi(\frac{7}{2}) \\ \phi(\frac{9}{2}) \\ \cdots \\ \phi(N-\frac{3}{2}) \\ \phi(N-\frac{1}{2}) \end{bmatrix} = \sqrt{2} \begin{bmatrix} \mathbf{c}(1) & \mathbf{c}(0) & & & & & & \\ \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & & \cdots & & \\ \mathbf{c}(5) & \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \cdots & & \\ \mathbf{c}(7) & \mathbf{c}(6) & \mathbf{c}(5) & \mathbf{c}(4) & \mathbf{c}(3) & \cdots & & \\ \mathbf{c}(9) & \mathbf{c}(8) & \mathbf{c}(7) & \mathbf{c}(6) & \mathbf{c}(5) & \cdots & & \\ & & & \cdots & & \cdots & & \\ & & & \cdots & & \cdots & \mathbf{c}(N-1) & \mathbf{c}(N-2) \\ & & & & & \cdots & 0 & \mathbf{c}(N) \end{bmatrix} \begin{bmatrix} \phi(0) \\ \phi(1) \\ \phi(2) \\ \phi(3) \\ \phi(4) \\ \cdots \\ \phi(N-2) \\ \phi(N-1) \end{bmatrix}$$

or

$$\Phi(\frac{1}{2}) = \mathbf{m}(1)\Phi(0). \tag{7.58}$$

We can continue in this way to compute $\phi(t)$ at all dyadic points. A general dyadic point will be of the form $t = n + s$ where $n = 0, 1, \cdots, N-1$ and $s < 1$ is of the form $s = \frac{k}{2^j}$, $k = 0, 1, \cdots, 2^j - 1$, $j = 1, 2, \cdots$. The N-rowed vector $\Phi(s)$ contains all the terms $\phi(n+s)$ whose fractional part is s:

$$\Phi(s) = \begin{bmatrix} \phi(s) \\ \phi(s+1) \\ \phi(s+2) \\ \phi(s+3) \\ \phi(s+4) \\ \cdots \\ \phi(s+N-2) \\ \phi(s+N-1) \end{bmatrix}$$

Example 7

$$\Phi(\frac{1}{4}) = \mathbf{m}(0)\Phi(\frac{1}{2}) \qquad \Phi(\frac{3}{4}) = \mathbf{m}(1)\Phi(\frac{1}{2}).$$

There are two possibilities, depending on whether the fractional dyadic s is $< \frac{1}{2}$ or $\geq \frac{1}{2}$. If $2s < 1$ then we can substitute $t = s, s+1, s+2, \cdots, s+N-1$, recursively, into the dilation equation and obtain the result

$$\Phi(s) = \mathbf{m}(0)\Phi(2s).$$

If $2s \geq 1$ then if we substitute $t = s, s+1, s+2, \cdots, s+N-1$, recursively, into the dilation equation we find that the lowest order term on the right-hand side is $\phi(2s-1)$ and that our result is

$$\Phi(s) = \mathbf{m}(1)\Phi(2s-1).$$

If we set $\Phi(s) = 0$ for $s < 0$ or $s \geq 1$ then we have the

Theorem 48 *The general vector recursion for evaluating the scaling function at dyadic points is*

$$\Phi(s) = \mathbf{m}(0)\Phi(2s) + \mathbf{m}(1)\Phi(2s-1). \tag{7.59}$$

From this recursion we can compute explicitly the value of ϕ at the dyadic point $t = n + s$. Indeed we can write s as a dyadic “decimal” $s = .s_1s_2s_3\cdots$ where

$$s = \sum_{j\geq 1} \frac{s_j}{2^j}, \qquad s_j = 0, 1.$$

If $2s < 1$ then $s_1 = 0$ and we have

$$\Phi(s) = \Phi(.0s_2s_3\cdots) = \mathbf{m}(0)\Phi(2s) = \mathbf{m}(0)\Phi(.s_2s_3s_4\cdots).$$

If on the other hand, $2s \geq 1$ then $s_1 = 1$ and we have

$$\Phi(s) = \Phi(.1s_2s_3\cdots) = \mathbf{m}(1)\Phi(2s-1) = \mathbf{m}(1)\Phi(.s_2s_3s_4\cdots).$$

Iterating this process we have the

Corollary 14

$$\Phi(.s_1s_2\cdots s_\ell) = \mathbf{m}(s_1)\mathbf{m}(s_2)\cdots\mathbf{m}(s_\ell)\Phi(0).$$

Note that the reasoning used to derive the recursion (7.59) for dyadic s also applies for a general real t such that $0 \le t < 1$. If we set $\Phi(t) = 0$ for $t < 0$ or $t \ge 1$ then we have the

Corollary 15

$$\Phi(t) = \mathbf{m}(0)\Phi(2t) + \mathbf{m}(1)\Phi(2t-1). \tag{7.60}$$

Note from (7.58) that the column sums of the $N \times N$ matrix $\mathbf{m}(1)$ are also 1 for each column, just as for the matrix $\mathbf{m}(0)$. Furthermore the column vector $\mathbf{e}_0 = (1, 1, \cdots, 1)$ is an eigenvector of $\mathbf{m}^{\rm tr}(0)$.

Denote by $< \Phi(t), \mathbf{e}_0 >= \sum_{n=0}^{N-1} \phi(t+n)$ the dot product of the vectors $\mathbf{e}_0$ and $\Phi(t)$. Taking the dot product of $\mathbf{e}_0$ with each side of the recursion (7.60) we find

$$< \Phi(t), \mathbf{e}_0 >=< \Phi(2t), \mathbf{e}_0 > + < \Phi(2t-1), \mathbf{e}_0 > .$$

If t is dyadic, we can follow the recursion backwards to $\Phi(0)$ and obtain the

Corollary 16

$$\sum_n \phi(s+n) = \sum_n \phi(n)$$

for all fractional dyadics s.

Thus the sum of the components of each of the dyadic vectors $\Phi(s)$ is constant. We can normalize this sum by requiring it to be 1, i.e., by requiring that $\sum_{n=0}^{N-1} \phi(n) = 1$. However, we have already normalized $\phi(t)$ by the requirement that $\int \phi(t)dt = 1$. Isn't there a conflict?

Not if $\phi(t)$ is obtained from the cascade algorithm. By substituting into the cascade algorithm one finds

$$\Phi^{(i+1)}(t) = \mathbf{m}(0)\Phi^{(i)}(2t) + \mathbf{m}(1)\Phi^{(i)}(2t-1),$$

and, taking the dot product of $\mathbf{e}_0$ with both sides of the equation we have

$$< \Phi^{(i+1)}(t), \mathbf{e}_0 >=< \Phi^{(i)}(2t), \mathbf{e}_0 > + < \Phi^{(i)}(2t-1), \mathbf{e}_0 >,$$

where only one of the terms on the right-hand side of this equation is nonzero. By continuing to work backwards we can relate the column sum for stage $i+1$ to the column sum for stage 0: $< \Phi^{(i+1)}(t), \mathbf{e}_0 >=< \Phi^{(0)}(\tau), \mathbf{e}_0 >$, for some τ such that $0 \le \tau < 1$.

At the initial stage we have $\phi^{(0)}(t) = 1$ for $0 \le t < 1$, and $\phi^{(0)}(t) = 0$ elsewhere, the box function, so $< \Phi^{(0)}(\tau), \mathbf{e}_0 >= 1$, and the sum is 1 as also is the area under the box function and the L^2 normalization of $\phi^{(0)}$. Thus the sum of the integral values of $\phi^{(i)}$ is preserved at each stage of the calculation, hence in the limit. We have proved the

Corollary 17 *If $\phi(t)$ is obtained as the limit of the cascade algorithm then*

$$\sum_n \phi(t+n) = 1$$

for all t.

NOTE: Strictly speaking we have proven the corollary only for $0 \leq t < 1$. However, if $t = m + r$ where m is an integer and $0 \leq r < 1$ then in the summand we can set $t + n = r + (m + n) = r + n'$ and sum over n' to get the desired result.

REMARK: We have shown that the $N \times N$ matrix $\mathbf{m}(0)$ has a column sum of 1, so that it always has the eigenvalue 1. If the eigenvalue 1 occurs with multiplicity one, then the matrix eigenvalue equation $\Phi(0) = \mathbf{m}(0)\Phi(0)$ will yield $N - 1$ linearly independent conditions for the N unknowns $\phi(0), \cdots, \phi(N-1)$. These together with the normalization condition $\sum_n \phi(t+n) = 1$ will allow us to solve (uniquely) for the N unknowns via Gaussian elimination. If $\mathbf{m}(0)$ has 1 as an eigenvalue with multiplicity $k > 1$ however, then the matrix eigenvalue equation will yield only $N - k$ linearly independent conditions for the N unknowns and these together with the normalization condition may not be sufficient to determine a unique solution for the N unknowns. A spectacular example (L^2 convergence of the scaling function, but it blows up at every dyadic point) is given at the bottom of page 248 in the text by Strang and Nguyen. The double eigenvalue 1 is not common, but not impossible.

Example 8 *The Daubechies filter coefficients for D_4 ($N = 3$) are $4\sqrt{2}\mathbf{c}(k) = 1+\sqrt{3}, 3+\sqrt{3}, 3-\sqrt{3}, 1-\sqrt{3}$. The equation*

$$\Phi(0) = \mathbf{m}(0)\Phi(0)$$

is in this case

$$\begin{bmatrix} \phi(0) \\ \phi(1) \\ \phi(2) \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1+\sqrt{3} & 0 & 0 \\ 3-\sqrt{3} & 3+\sqrt{3} & 1+\sqrt{3} \\ 0 & 1-\sqrt{3} & 3-\sqrt{3} \end{bmatrix} \begin{bmatrix} \phi(0) \\ \phi(1) \\ \phi(2) \end{bmatrix}.$$

Thus with the normalization $\phi(0) + \phi(1) + \phi(2) = 1$ we have, uniquely,

$$\phi(0) = 0, \quad \phi(1) = \frac{1}{2}(1+\sqrt{3}) \quad \phi(2) = \frac{1}{2}(1-\sqrt{3}).$$

REMARK 1: The preceding corollary tells us that, locally at least, we can represent constants within the multiresolution space V_0. This is related to the fact that $\mathbf{e}$ is a left eigenvector for $\mathbf{m}(0)$ and $\mathbf{m}(1)$ which is in turn related to the fact that $C(\omega)$ has a zero at $\omega = \pi$. We will show later that if we require that $C(\omega)$ has a zero of order p at $\omega = \pi$ then we will be able to represent the monomials $1, t, t^2, \cdots t^{p-1}$ within V_0, hence all polynomials in t of order $p-1$. This is a highly desirable feature for wavelets and is satisfied by the Daubechies wavelets of order p.

REMARK 2: In analogy with the use of infinite matrices in filter theory, we can also relate the dilation equation

$$\phi(t) = \sqrt{2}\sum_{k=0}^{N} \mathbf{c}(k)\phi(2t-k)$$

to an infinite matrix $\mathbf{M}$. Evaluate the equation at the values $t+n, n = 0, \pm 1, \cdots$ for any real t. Substituting these values one at a time into the dilation equation we obtain the system of equations

$$\begin{bmatrix} \vdots \\ \phi(t-2) \\ \phi(t-1) \\ \phi(t) \\ \phi(t+1) \\ \phi(t+2) \\ \vdots \end{bmatrix} = \sqrt{2}\begin{bmatrix} \cdots & & & \cdots & & & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & \mathbf{c}(0) & 0 & 0 & 0 & 0 & \cdots \\ \cdots & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & 0 & 0 & \cdots \\ \cdots & \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \mathbf{c}(1) & \mathbf{c}(0) & \cdots \\ \cdots & \mathbf{c}(6) & \mathbf{c}(5) & \mathbf{c}(4) & \mathbf{c}(3) & \mathbf{c}(2) & \cdots \\ \cdots & & & \cdots & & & \cdots \end{bmatrix}\begin{bmatrix} \vdots \\ \phi(2t-2) \\ \phi(2t-1) \\ \phi(2t) \\ \phi(2t+1) \\ \phi(2t+2) \\ \vdots \end{bmatrix},$$

or

$$\Phi_\infty(t) = \mathbf{M}\Phi_\infty(2t), \qquad -\infty < t < \infty. \tag{7.61}$$

We have met $\mathbf{M}$ before. The matrix elements of $\mathbf{M}$ are $\mathbf{M}_{ij} = \sqrt{2}\mathbf{c}(2i-j)$. Note the characteristic double-shift of the rows. We have

$$\mathbf{M} = \sqrt{2}\mathbf{L} = (\downarrow 2)\sqrt{2}\mathbf{C}$$

where $\mathbf{L}$ is the double-shifted matrix corresponding to the low pass filter $\mathbf{C}$. For any fixed t the only nonzero part of (7.61) will correspond to either $\mathbf{m}(0)$ or $\mathbf{m}(1)$. For $0 \le t < 1$ the equation reduces to (7.60). $\mathbf{M}$ shares with its finite forms $\mathbf{m}(i)$ the fact that the column sum is 1 for every column and that $\lambda = 1$ is an eigenvalue. The left eigenvector now, however, is the infinite component row vector $\mathbf{e}^{\infty}{}_0 = (\cdots, 1, 1, 1, \cdots)$. We take the inner product of this vector only with other vectors that are finitely supported, so there is no convergence problem.

Now we are in a position to investigate some of the implications of requiring that $C(\omega)$ has a zero of order p at $\omega = \pi$ for $p > 1$ This requirement means that

$$C(\pi) = C'(\pi) = \cdots = C^{(p-1)}(\pi) = 0,$$

and since $C(\omega) = \sum_k \mathbf{c}(k)e^{-ik\omega}$, it is equivalent to

$$\sum_k (-1)^k \mathbf{c}(k) k^\ell = 0, \qquad \ell = 1, 2, \cdots, p-1. \tag{7.62}$$

We already know that

$$\sum_k \mathbf{c}(2k) = \sum_k \mathbf{c}(2k+1) = \frac{1}{\sqrt{2}} = \frac{1}{2} C(0), \tag{7.63}$$

and that $|C(\omega)|^2 + |C(\omega+\pi)|^2 = 2$. For use in the proof of the theorem to follow, we introduce the notation

$$A_\ell = \sum_i (2i)^\ell \mathbf{c}(2i) = \sum_i (2i+1)^\ell \mathbf{c}(2i+1), \quad \ell = 0, 1, \cdots, p-1. \tag{7.64}$$

We already know that $A_0 = \frac{1}{\sqrt{2}}$.

The condition that $\mathbf{M}$ admit a left eigenvector $\mathbf{e} = (\cdots, \alpha_{-1}, \alpha_0, \alpha_1, \cdots)$ with eigenvalue λ is that the equations

$$\sqrt{2} \sum_i \alpha_i \mathbf{c}(2i-j) = \lambda \alpha_j, \qquad i, j = 0, \pm 1, \cdots \tag{7.65}$$

hold where not all α_i are zero. A similar statement holds for the finite matrices $\mathbf{m}(0), \mathbf{m}(1)$ except that i, j are restricted to the rows and columns of these finite matrices. (Indeed the finite matrices $\mathbf{m}(0)_{ij} = \sqrt{2}\mathbf{c}(2i-j)$ for $0 \le i, j \le N-1$ and $\mathbf{m}(1)_{ij} = \sqrt{2}\mathbf{c}(2i-j+1)$ for $0 \le i, j \le N-1$ have the property that the jth column vector of $\mathbf{m}(0)$ and the $(j+1)$st column vector of $\mathbf{m}(1)$ each contain all of the nonzero elements in the jth column of the infinite matrix $\mathbf{M}$. Thus the restriction of (7.65) to the row and column indices i, j for $\mathbf{m}(0)$, $\mathbf{m}(1)$ yields exactly the eigenvalue equations for these finite matrices.) We have already shown that this equation has the solution $\alpha_i = 1, \lambda = 1$, due to that fact that $C(\omega)$ has a zero of order 1 at π.

For each integer h we define the (infinity-tuple) row vector $\mathbf{e_h}^\infty$ by

$$(\mathbf{e_h}^\infty)_i = i^h, \qquad i = 0, \pm 1 \cdots.$$

Theorem 49 *If $C(\omega)$ has a zero of order $p > 1$ at $\omega = \pi$ then $\mathbf{M}$ (and $\mathbf{m}(0), \mathbf{m}(1)$) have eigenvalues $\lambda_\ell = \frac{1}{2^\ell}, \quad \ell = 0, 1, \cdots, p-1$. The corresponding left eigenvectors $\mathbf{y}_\ell$ can be expressed as*

$$(-1)^\ell \mathbf{y}_\ell = \mathbf{e}_\ell{}^\infty + \sum_{h=0}^{\ell-1} \alpha_h^\ell \mathbf{e_h}{}^\infty.$$

PROOF: For each $\ell = 0, 1, \cdots, p-1$ we have to verify that an identity of the following form holds:

$$\sqrt{2}\sum_i \left(i^\ell + \sum_{h=0}^{\ell-1} \alpha_h^\ell i^h \right) \mathbf{c}(2i - j) = \frac{1}{2^\ell} \left(j^\ell + \sum_{h=0}^{\ell-1} \alpha_h^\ell j^h \right),$$

for $i, j = 0, \pm 1, \pm 2, \cdots$. For $\ell = 0$ we already know this. Suppose $\ell \geq 1$.

Take first the case where $j = 2s$ is even. We must find constants α_h^ℓ such that the identity

$$\sqrt{2}\sum_i \left(i^\ell + \sum_{h=0}^{\ell-1} \alpha_h^\ell i^h \right) \mathbf{c}(2i - 2s) = \frac{1}{2^\ell} \left((2s)^\ell + \sum_{h=0}^{\ell-1} \alpha_h^\ell (2s)^h \right),$$

holds for all s. Making the change of variable $i' = i - s$ on the left-hand side of this expression we obtain

$$\sqrt{2}\sum_{i'} \left((i'+s)^\ell + \sum_{h=0}^{\ell-1} \alpha_h^\ell (i'+s)^h \right) \mathbf{c}(2i') = \frac{1}{2^\ell} \left((2s)^\ell + \sum_{h=0}^{\ell-1} \alpha_h^\ell (2s)^h \right).$$

Expanding the left-hand side via the binomial theorem and using the sums (7.64) we find

$$\sqrt{2}\sum_n s^n \left[\left(\begin{array}{c} \ell \\ n \end{array} \right) \frac{A_{\ell-n}}{2^{\ell-n}} + \sum_{h=0}^{\ell-1} \alpha_h^\ell \left(\begin{array}{c} h \\ n \end{array} \right) \frac{A_{h-n}}{2^{h-n}} \right]$$
$$= \frac{1}{2^\ell} \left((2s)^\ell + \sum_{h=0}^{\ell-1} \alpha_h^\ell (2s)^h \right).$$

Now we equate powers of s. The coefficient of s^ℓ on both sides is 1. Equating coefficients of $s^{\ell-1}$ we find

$$\frac{\ell A_1}{\sqrt{2}} + \alpha_{\ell-1}^\ell = \frac{1}{2}\alpha_{\ell-1}^\ell.$$

We can solve for $\alpha^\ell_{\ell-1}$ in terms of the given sum A_1. Now the pattern becomes clear. We can solve these equations recursively for $\alpha^\ell_{\ell-1}, \alpha^\ell_{\ell-2}, \cdots \alpha^\ell_0$. Equating coefficients of $s^{\ell-j}$ allows us to express $\alpha^\ell_{\ell-j}$ as a linear combination of the A_i and of $\alpha^\ell_{\ell-1}, \alpha^\ell_{\ell-2}, \cdots \alpha^\ell_{j+1}$. Indeed the equation for $\alpha^\ell_{\ell-j}$ is

$$\alpha^\ell_{\ell-j} = \frac{\sqrt{2}}{2^{-j}-1}\left[\begin{pmatrix} \ell \\ \ell-j \end{pmatrix}\frac{A_j}{2^j} + \sum_{h=\ell-j+1}^{\ell-1} \alpha^\ell_h \begin{pmatrix} h \\ n \end{pmatrix} \frac{A_{h-\ell+j}}{2^{h-\ell+j}}\right].$$

This finishes the proof for $j = 2s$. The proof for $j = 2s-1$ follows immediately from replacing s by $s - \frac{1}{2}$ in our computation above and using the fact that the A_h are the same for the sums over the even terms in **c** as for the sums over the odd terms. Q.E.D.

Since $\Phi_\infty(t) = \mathbf{M}\Phi_\infty(2t)$ and $\mathbf{y}_\ell \cdot \mathbf{M} = \frac{1}{2^\ell}\mathbf{y}_\ell$ for $\ell = 1, \cdots, p-1$ it follows that the function

$$\mathbf{y}_\ell \cdot \Phi_\infty(t) = \sum_k (\mathbf{y}_\ell)_i \phi(t+k)$$

satisfies

$$\mathbf{y}_\ell \cdot \Phi_\infty(t) = \frac{1}{2^\ell}\mathbf{y}_\ell \cdot \Phi_\infty(2t).$$

Iterating this identity we have

$$\mathbf{y}_\ell \cdot \Phi_\infty(t) = \frac{1}{2^{n\ell}}\mathbf{y}_\ell \cdot \Phi_\infty(2^n t)$$

for $n = 1, 2, \cdots$. Hence

$$\mathbf{y}_\ell \cdot \Phi_\infty(t) = \lim_{n\to+\infty} \frac{1}{2^{n\ell}}\mathbf{y}_\ell \cdot \Phi_\infty(2^n t). \tag{7.66}$$

We can compute this limit explicitly. Fix $t_0 \neq 0$ and let n be a positive integer. Denote by $[2^n t_0]$ the largest integer $\leq 2^n t_0$. Then $2^n t_0 = [2^n t_0] + s_n$ where $0 \leq s_n < 1$. Since the support of $\phi(t)$ is contained in the interval $[0, N]$, the only nonzero terms in $\mathbf{y}_\ell \cdot \Phi_\infty(2^n t_0)$ are

$$\mathbf{y}_\ell \cdot \Phi_\infty(2^n t_0) = \sum_{j=0}^{N-1} (\mathbf{y}_\ell)_{j-[2^n t_0]}\phi(s_n + j).$$

Now as $n \to +\infty$ the terms s_n all lie in the range $0 \leq s_n < 1$ so we can find a subsequence $\{s_{n_h} : h = 1, 2, \cdots\}$ such that the subsequence converges to $s \in [0, 1]$,

$$\lim_{h\to+\infty} s_{n_h} = s.$$

Since $\phi(t)$ is continuous we have $\phi(s_{n_h}+j)\rightarrow\phi(s+j)$ as $h\rightarrow\infty$. Further, since $(\mathbf{y}_\ell)_i=(-i)^\ell+$ lower order terms in i we see that

$$\lim_{h\rightarrow+\infty}\frac{1}{2^{n_h\ell}}\sum_{j=0}^{N-1}(\mathbf{y}_\ell)_{j-[2^n_h t_0]}\phi(s_{n_h}+j)=(t_0)^\ell\sum_{j=0}^{N-1}\phi(s+j)=(t_0)^\ell,$$

since always $\sum_{j=0}^{N-1}\phi(s+j)=1$. If $t_0=0$ then the limit expression (7.66) shows directly that $\mathbf{y}_\ell\cdot\Phi_\infty(0)=0$. Thus we have the

Theorem 50 *If $H(\omega)$ has $p\geq 1$ zeros at $\omega=\pi$ then*

$$\sum_k \mathbf{y}_{\ell k}\phi(t+k)=t^\ell,\qquad \ell=0,1,\cdots,p-1$$

The result that a bounded sequence of real numbers contains a convergent subsequence is standard in analysis courses. For completeness, I will give a proof, tailored to the problem at hand.

Lemma 44 *Let $\{x_n:n=1,2,\cdots\}$ be sequence of real numbers in the bounded interval $[0,1]$: $0\leq x_n\leq 1$. There exists a convergent subsequence $\{x_{n_h}:h=1,2,\cdots\}$, i.e., $\lim_{h\rightarrow\infty}x_{n_h}=s$, where $0\leq s\leq 1$.*

PROOF: Consider the dyadic representation for a real number x on the unit interval $[0,1]$:

$$x=.s_1s_2s_3\cdots s_j\cdots=\sum_{j=1}^{\infty}\frac{s_j}{2^j},$$

where $s_j=0,1$. Our given sequence $\{x_n\}$ contains a countably infinite number of elements, not necessarily distinct. If the subinterval $[0,1/2)$ contains a countably infinite number of elements x_n, choose one x_{n_1}, set $s_1=0$ and consider only the elements $x_n\in[0/1/2)\equiv I_1$ with $n>n_1$. If the subinterval $[0,1/2)$ contains only finitely many elements x_n, choose $x_{n_1}\in[1/2,1]$, set $s_1=1$, and consider only the elements $x_n\in[1/2,1]\equiv I_1$ with $n>n_1$. Now repeat the process in the interval I_2, dividing it into two subintervals of length $1/2^2$ and setting $s_2=0$ if there are an infinite number of elements of the remaining sequence in the left-hand subinterval; or $s_2=1$ if there are not, and choosing x_{n_2} from the first infinite interval. Continuing this way we obtain a sequence of numbers $s_1,s_2,\cdots$ where $s_h=0,1$ and a subsequence $\{x_{n_h}\}$ such that $\lim_{h\rightarrow\infty}x_{n_h}=s$ where $s=.s_1s_2\cdots$ in dyadic notation. Q.E.D.

Theorem 50 shows that if we require that $C(\omega)$ has a zero of order p at $\omega = \pi$ then we can represent the monomials $1, t, t^2, \cdots t^{p-1}$ within V_0, hence all polynomials in t of order $p-1$. This isn't quite correct since the functions $\mathbf{y}_\ell \cdot \Phi_\infty(t) = t^\ell$, strictly speaking, don't belong to V_0, or even to $L^2[-\infty, \infty]$. However, due to the compact support of the scaling function, the series converges pointwise. Normally one needs only to represent the polynomial in a bounded domain. Then all of the coefficients $\mathbf{y}_\ell$ that don't contribute to the sum in that bounded interval can be set equal to zero.

7.6 Infinite product formula for the scaling function

We have been studying pointwise convergence of iterations of the dilation equation in the time domain. Now we look at the dilation equation in the frequency domain. The equation is

$$\phi(t) = 2\sum_k \mathbf{h}(k)\phi(2t-k),$$

where $\mathbf{c}(k) = \sqrt{2}\mathbf{h}(k)$. Taking the Fourier transform of both sides of this equation and using the fact that (for $u = 2t - k$)

$$2\int_{-\infty}^{\infty} \phi(2t-k)e^{-i\omega t}dt = \int_{-\infty}^{\infty} \phi(u)e^{-i\omega(u+k)/2}du = e^{-i\omega k/2}\hat{\phi}(\frac{\omega}{2}).$$

we find

$$\hat{\phi}(\omega) = \left(\sum_k \mathbf{h}(k)e^{-i\omega k/2}\right)\hat{\phi}(\frac{\omega}{2}).$$

Thus the frequency domain form of the dilation equation is

$$\hat{\phi}(\omega) = H(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2}). \tag{7.67}$$

(Here $C(\omega) = \sqrt{2}H(\omega)$. We have changed or normalization because the property $H(0) = 1$ is very convenient for the cascade algorithm.) Now iterate the right-hand side of the dilation equation:

$$\hat{\phi}(\omega) = H(\frac{\omega}{2})[H(\frac{\omega}{4})\hat{\phi}(\frac{\omega}{4})].$$

After N steps we have

$$\hat{\phi}(\omega) = H(\frac{\omega}{2})H(\frac{\omega}{4})\cdots H(\frac{\omega}{2^N})\hat{\phi}(\frac{\omega}{2^N}).$$

We want to let $N \to \infty$ on the right-hand side of this equation. Proceeding formally, we note that if the cascade algorithm converges it will yield a scaling function $\phi(t)$ such that $\hat{\phi}(0) = \int \phi(t)dt = 1$. Thus we assume $\lim_{N\to\infty} \hat{\phi}(\frac{\omega}{2^N}) = \hat{\phi}(0) = 1$ and postulate an infinite product formula for $\hat{\phi}(\omega)$:

$$\hat{\phi}(\omega) = \Pi_{j=1}^{\infty} H(\frac{\omega}{2^j}). \tag{7.68}$$

TIME OUT: **Some facts about the pointwise convergence of infinite products.**

An infinite product is usually written as an expression of the form

$$P \equiv \Pi_{j=1}^{\infty}(1 + w_j), \tag{7.69}$$

where $\{w_j\}$ is a sequence of complex numbers. (In our case $1 + w_j = H(\frac{\omega}{2^2})$). An obvious way to attempt to define precisely what it means for an infinite product to converge is to consider the finite products

$$P_n = \Pi_{j=1}^{n}(1 + w_j)$$

and say that the infinite product is convergent if $\lim_{n\to\infty} P_n$ exists as a finite number. However, this is a bit too simple because if $1 + w_m = 0$ for any term m, then the product will be zero regardless of the behavior of the rest of the terms. What we do is to allow a *finite* number of the factors to vanish and then require that if these factors are omitted then the remaining infinite product converges in the sense stated above. Thus we have the

Definition 31 *Let*

$$P_{m,n} = \Pi_{j=m}^{n}(1 + w_j), \qquad 1 \le m < n.$$

The infinite product (7.69) is convergent if

1. *there exists an $m_0 \ge 1$ such that $w_j \ne -1$ for $k \ge m_0$, and*
2. *for $m \ge m_0$*
$$\lim_{n\to\infty} P_{m,n}$$
exists as a finite nonzero number.

Thus the Cauchy criterion for convergence is, given any $\epsilon > 0$ there must exist an $N(\epsilon)$ such that $|P_{m,n} - P_{m,n+p}| = |P_{m,n}| \cdot |1 - P_{n+1,n+p}| < \epsilon$ for all $m \ge m_0$, $n > N$ and $p \ge 1$.

The basic convergence tool is the following:

Theorem 51 *If $p_j \geq 0$ for all j, then the infinite product*

$$\Pi_{j=1}^{\infty}(1 + p_j)$$

converges if and only if $\sum_{j=1}^{\infty} p_j$ converges.

PROOF: Set

$$\mathcal{P}_n = \sum_{j=1}^{n} p_j, \qquad P_n = \Pi_{j=1}^{n}(1 + p_j).$$

Now

$$1 + \mathcal{P}_n \leq P_n = \Pi_{j=1}^{n}(1 + p_j) \leq e^{\sum_{j=1}^{n} p_j} = e^{\mathcal{P}_n},$$

where the last inequality follows from $1 + p_j \leq e^{p_j}$. (The left-hand side is just the first two terms in the power series of e^{p_j}, and the power series contains only nonnegative terms.) Thus the infinite product converges if and only if the power series converges. Q.E.D.

Definition 32 *We say that an infinite product is absolutely convergent if the infinite product $\Pi_{j=1}^{\infty}(1 + |w_j|)$ is convergent.*

Theorem 52 *An absolutely convergent infinite product is convergent.*

PROOF: If the infinite product P is absolutely convergent, then $Q \equiv \Pi_{j=1}^{\infty}(1 + |w_j|)$ is convergent and $\sum_j |w_j| < \infty$, so $w_j \to 0$. It is a simple matter to check that each of the factors in the expression $|P_{m,n}| \cdot |1 - P_{n+1,n+p}|$ is bounded above by the corresponding factor in the convergent product Q. Hence the sequence is Cauchy and P is convergent. Q.E.D.

Definition 33 *Let $\{w_j(z)\}$ be a sequence of continuous functions defined on an open connected set D of the complex plane, and let S be a closed, bounded subset of D. The infinite product $P(z) \equiv \Pi_{j=1}^{\infty}(1 + w_j(z))$ is said to be uniformly convergent on S if*

1. *there exists a fixed $m_0 \geq 1$ such that $w_j(z) \neq -1$ for $k \geq m_0$, and every $z \in S$, and*

2. *for any $\epsilon > 0$ there exists a fixed $N(\epsilon)$ such that for $n > N(\epsilon)$, $m \geq m_0$ and every $z \in S$ we have*

$$|P_{m,n}(z)| \cdot |1 - P_{n+1,n+p}(z)| < \epsilon, p \geq 1.$$

Then from standard results in calculus we have the

Theorem 53 *Suppose $w_j(z)$ is continuous in D for each j and that the infinite product $P(z) \equiv \Pi_{j=1}^{\infty}(1 + w_j(z))$ converges uniformly on every closed bounded subset of D. Then $P(z)$ is a continuous function in D.*

BACK TO THE INFINITE PRODUCT FORMULA FOR THE SCALING FUNCTION:

$$\hat{\phi}(\omega) = \Pi_{j=1}^{\infty} H(\frac{\omega}{2^j}).$$

Note that this infinite product converges, uniformly and absolutely on all finite intervals. Indeed note that $H(0) = 1$ and that the derivative of the 2π-periodic function $H'(\omega)$ is uniformly bounded: $|H'(\omega)| \leq C$. Then $H(\omega) = H(0) + \int_0^{\omega} H'(s)ds$ so

$$|H(\omega)| \leq 1 + C|\omega| \leq e^{C|\omega|}.$$

Since $C\sum_{j=1}^{\infty} \frac{|\omega|}{2^j} = C|\omega|$ converges, the infinite product converges absolutely, and we have the (very crude) upper bound $|\hat{\phi}(\omega)| \leq e^{C|\omega|}$.

Example 9 *The moving average filter has filter coefficients $\mathbf{h}(0) = \mathbf{h}(1) = \frac{1}{2}$ and $H(\omega) = \frac{1}{2}(1 + e^{-i\omega})$. The product of the first N factors in the infinite product formula is*

$$H^{(N)}(\omega) = \frac{1}{2^N}(1 + e^{-i\omega/2})(1 + e^{-i\omega/4})(1 + e^{-i\omega/8}) \cdots (1 + e^{-i\omega/2^N}).$$

The following identities (easily proved by induction) are needed:

Lemma 45

$$(1 + z)(1 + z^2)(1 + z^4) \cdots (1 + z^{2^{n-1}}) = \sum_{k=0}^{2^n-1} z^k = \frac{1 - z^{2^n}}{1 - z}.$$

Then, setting $z = e^{-i\omega/2^N}$, we have

$$H^{(N)}(\omega) = \frac{1}{2^N}\frac{1 - e^{-i\omega}}{1 - e^{-i\omega/2^N}}.$$

Now let $N \to \infty$. The numerator is constant. The denominator goes like $2^N(i\omega/2^N - \omega^2/2^{2N+1} + \cdots) \to i\omega$. Thus

$$\hat{\phi}(\omega) = \lim_{N\to\infty} H^{(N)}(\omega) = \frac{1 - e^{-i\omega}}{i\omega}, \qquad (7.70)$$

basically the sinc function.

Although the infinite product formula for $\hat{\phi}(\omega)$ always converges pointwise and uniformly on any closed bounded interval, this doesn't solve our problem. We need to have $\hat{\phi}(\omega)$ decays sufficiently rapidly at infinity so that it belongs to L^2. At this point all we have is a weak solution to the problem. The corresponding $\phi(t)$ is not a function but a *generalized function* or *distribution*. One can get meaningful results only in terms of integrals of $\phi(t)$ and $\hat{\phi}(\omega)$ with functions that decay very rapidly at infinity and their Fourier transforms that also decay rapidly. Thus we can make sense of the generalized function $\phi(t)$ by defining the expression on the left-hand side of

$$2\pi \int_{-\infty}^{\infty} \phi(t)\overline{g}(t)dt = \int_{-\infty}^{\infty} \hat{\phi}(\omega)\overline{\hat{g}}(\omega)d\omega$$

by the integral on the right-hand side, for all g and $\hat{g}$ that decay sufficiently rapidly. We shall not go that route because we want $\phi(t)$ to be a true function.

Already the crude estimate $|\hat{\phi}(\omega)| < e^{C|\omega|}$ in the complex plane does give us some information. The Paley-Weiner Theorem (whose proof is beyond the scope of this course) says, essentially that for a function $\phi(t) \in L^2[-\infty, \infty]$ the Fourier transform can be extended into the complex plane such that $|\hat{\phi}(\omega)| < Ke^{C|\omega|}$ if and only if $\phi(t) \equiv 0$ for $|t| > C$. It is easy to understand why this is true. If $\phi(t)$ vanishes for $|t| > C$ then $\hat{\phi}(\omega) = \int_{-C}^{C} \phi(t)e^{-i\omega t}dt$ can be extended into the complex ω plane and the above integral satisfies this estimate. If $\phi(t)$ is nonzero in an interval around t_0 then it will make a contribute to the integral whose absolute value would grow at the approximate rate $e^{|t_0\omega|}$.

Thus we know that if $\hat{\phi}$ belongs to L^2, so that $\phi(t)$ exists, then $\phi(t)$ has compact support. We also know that if $\sum_{k=0}^{N} \mathbf{h}(k) = 1$, our solution (if it exists) is unique.

Let's look at the shift orthogonality of the scaling function in the frequency domain. Following Strang and Nguyen we consider the inner product vector

$$\mathbf{a}(k) = (\phi_{00}, \phi_{0k}) = \int_{-\infty}^{\infty} \phi(t)\overline{\phi}(t-k)dt, \qquad (7.71)$$

and its associated finite Fourier transform $A(\omega) = \sum_k \mathbf{a}(k)e^{-ik\omega}$. Note that the integer translates of the scaling function are orthonormal if and only if $\mathbf{a}(k) = \delta_{0k}$, i.e., $A(\omega) \equiv 1$. However, for later use in the study of biorthogonal wavelets, we shall also consider the possibility that the translates are not orthonormal.

Using the Plancherel equality and the fact that the Fourier transform of $\phi(t-k)$ is $e^{-ik\omega}\hat{\phi}(\omega)$ we have in the frequency domain

$$\mathbf{a}(k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{\phi}(\omega)\overline{\hat{\phi}}(\omega)e^{ik\omega}d\omega$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \sum_{n=-\infty}^{\infty} |\hat{\phi}(\omega+2\pi n)|^2 e^{ik\omega} d\omega.$$

Theorem 54

$$A(\omega) = \sum_{-\infty}^{\infty} |\hat{\phi}(\omega+2\pi n)|^2.$$

The integer translates of $\phi(t)$ are orthonormal if and only if $A(\omega) \equiv 1$.

The function $A(\omega)$ and its transform, the vector of inner products $(\phi(t), \phi(t+k)$ will be major players in our study of the L^2 convergence of the cascade algorithm. Let's derive some of its properties with respect to the dilation equation. We will express $A(2\omega)$ in terms of $A(\omega)$ and $H(\omega)$. Since $\hat{\phi}(2\omega) = H(\omega)\hat{\phi}(\omega)$ from the dilation equation, we have

$$\hat{\phi}(2\omega+2\pi n) = H(\omega+\pi n)\hat{\phi}(\omega+\pi n)$$

$$= \begin{cases} H(\omega)\hat{\phi}(\omega+2\pi k) & n = 2k \\ H(\omega+\pi)\hat{\phi}(\omega+\pi+2\pi k) & n = 2k+1. \end{cases}$$

since $H(\omega+2\pi) = H(\omega)$. Squaring and adding to get $A(2\omega)$ we find

$$\begin{aligned} A(2\omega) &= |H(\omega)|^2 \sum_k |\hat{\phi}(\omega+2\pi k)|^2 + |H(\omega+\pi)|^2 \sum_k |\hat{\phi}(\omega+\pi+2\pi k)|^2 \\ &= |H(\omega)|^2 A(\omega) + |H(\omega+\pi)|^2 A(\omega+\pi) \qquad (7.72) \end{aligned}$$

Essentially the same derivation shows how $A(\omega)$ changes with each pass through the cascade algorithm. Let

$$\mathbf{a}^{(i)}(k) = (\phi^{(i)}_{00}, \phi^{(i)}_{0k}) = \int_{-\infty}^{\infty} \phi^{(i)}(t)\overline{\phi}^{(i)}(t-k)dt, \qquad (7.73)$$

and its associated Fourier transform $A^{(i)}(\omega) = \sum_k \mathbf{a}^{(i)}(k)e^{-ik\omega}$ denote the information about the inner products of the functions $\phi^{(i)}(t)$ obtained from the ith passage through the cascade algorithm. Since $\hat{\phi}^{(i+1)}(2\omega) = H(\omega)\hat{\phi}^{(i)}(\omega)$ we see immediately that

$$A^{(i+1)}(2\omega) = |H(\omega)|^2 A^{(i)}(\omega) + |H(\omega+\pi)|^2 A^{(i)}(\omega+\pi) \qquad (7.74)$$

Chapter 8

Wavelet Theory

In this chapter we will provide some solutions to the questions of the existence of wavelets with compact and continuous scaling functions, the L^2 convergence of the cascade algorithm and the accuracy of approximation of functions by wavelets.

At the end of the preceding chapter we introduced, in the frequency domain, the transformation that relates the inner products

$$\int_{-\infty}^{\infty} \phi^{(i+1)}(t)\overline{\phi}^{(i+1)}(t-k)dt$$

to the inner products $\int_{-\infty}^{\infty} \phi^{(i)}(t)\overline{\phi}^{(i)}(t-k)dt$ in successive passages through the cascade algorithm. In the time domain this relationship is as follows. Let

$$\mathbf{a}^{(i)}(k) = (\phi^{(i)}_{00}, \phi^{(i)}_{0k}) = \int_{-\infty}^{\infty} \phi^{(i)}(t)\overline{\phi}^{(i)}(t-k)dt, \tag{8.1}$$

be the vector of inner products at stage i. Note that although $\mathbf{a}^{(i)}$ is an infinite-component vector, since $\phi^{(i)}(t)$ has support limited to the interval $[0, N]$ only the $2N-1$ components $\mathbf{a}^{(i)}(k)$, $k = -N+1, \cdots -1, 0, 1 \cdots, N-1$ can be nonzero. We can use the cascade recursion to express $\mathbf{a}^{(i+1)}(s)$ as a linear combination of terms $\mathbf{a}^{(i)}(k)$:

$$\mathbf{a}^{(i+1)}(s) = \int_{-\infty}^{\infty} \overline{\phi}^{(i+1)}(t)\phi^{(i+1)}(t+s)dt$$

$$= 4\sum_{k,\ell} \overline{\mathbf{h}}(k)\mathbf{h}(\ell) \int_{-\infty}^{\infty} \overline{\phi}^{(i)}(t-k)\phi^{(i)}(2t+2s-\ell)dt$$

$$= 2\sum_{j,\ell} \mathbf{h}(2s-j)\overline{\mathbf{h}}(\ell-j) \int_{-\infty}^{\infty} \overline{\phi}^{(i)}(t)\phi^{(i)}(t+\ell)dt.$$

Thus

$$\mathbf{a}^{(i+1)}(s) = 2\sum_{j,\ell} \mathbf{h}(2s+j)\overline{\mathbf{h}}(\ell+j)\mathbf{a}^{(i)}(\ell) \tag{8.2}$$

(Recall that we have already used this same recursion in the proof of Theorem 47.) In matrix notation this is just

$$\mathbf{a}^{(i+1)} = \mathbf{T}\mathbf{a}^{(i)} = (\downarrow 2)2\mathbf{H}\overline{\mathbf{H}}^{\mathrm{tr}}\mathbf{a}^{(i)} \tag{8.3}$$

where the matrix elements of the $\mathbf{T}$ matrix (the *transition matrix*) are given by

$$\mathbf{T}_{s\ell} = 2\sum_{j} \mathbf{h}(2s+j)\overline{\mathbf{h}}(\ell+j).$$

Although $\mathbf{T}$ is an infinite matrix, the only elements that correspond to inner products of functions with support in $[0, N]$ are contained in the $(2N-1)\times(2N-1)$ block $-N+1 \le s, \ell \le N-1$. When we discuss the eigenvalues and eigenvectors of $\mathbf{T}$ we are normally talking about this $(2N-1)\times(2N-1)$ matrix. We emphasis the relation with other matrices that we have studied before:

$$\mathbf{T} = (\downarrow 2)2\mathbf{H}\overline{\mathbf{H}}^{\mathrm{tr}} = \mathbf{M}\overline{\mathbf{H}}^{\mathrm{tr}}.$$

Note: Since The filter $\mathbf{H}$ is low pass, the matrix $\mathbf{T}$ shares with the matrix $\mathbf{M}$ the property that the column sum of each column equals 1. Indeed $\sum_s \mathbf{h}(2s+j) = \frac{1}{2}$ for all j and $\sum_j \mathbf{h}(j) = 1$, so $\sum_s \mathbf{T}_{s\ell} = 1$ for all ℓ. Thus, just as is the case with $\mathbf{M}$, we see that $\mathbf{T}$ admits the left-hand eigenvector $\mathbf{e}_0 = (\cdots, 1, 1, 1, \cdots)$ with eigenvalue 1.

If we apply the cascade algorithm to the inner product vector of the scaling function itself

$$\mathbf{a}(k) = (\phi_{00}, \phi_{0k}) = \int_{-\infty}^{\infty} \phi(t)\overline{\phi}(t-k)dt,$$

we just reproduce the inner product vector:

$$\mathbf{a}(s) = 2\sum_{j,\ell} \mathbf{h}(2s-j)\overline{\mathbf{h}}(\ell-j)\mathbf{a}(\ell), \tag{8.4}$$

or

$$\mathbf{a} = \mathbf{T}\mathbf{a} = (\downarrow 2)2\mathbf{H}\overline{\mathbf{H}}^{\mathrm{tr}}\mathbf{a} \tag{8.5}$$

Since $\mathbf{a}(k) = \delta_{0k}$ in the orthogonal case, this just says that

$$1 = 2\sum_{j} |\mathbf{h}(j)|^2,$$

which we already know to be true. Thus $\mathbf{T}$ always has 1 as an eigenvalue, with associated eigenvector $\mathbf{a}(k) = \delta_{0k}$.

If we apply the cascade algorithm to the inner product vector of the ith iterate of the cascade algorithm with the scaling function itself

$$\mathbf{b}(k) = \int_{-\infty}^{\infty} \phi(t)\overline{\phi}^{(i)}(t-k)dt,$$

we find, by the usual computation,

$$\mathbf{b}^{(i+1)} = \mathbf{T}\mathbf{b}^{(i)}. \qquad (8.6)$$

8.1 L^2 convergence

Now we have reached an important point in the convergence theory for wavelets! We will show that the necessary and sufficient condition for the cascade algorithm to converge in L^2 to a unique solution of the dilation equation is that the transition matrix $\mathbf{T}$ has a non-repeated eigenvalue 1 and all other eigenvalues λ such that $|\lambda| < 1$. Since the only nonzero part of $\mathbf{T}$ is a $(2N-1)\times(2N-1)$ block with very special structure, this is something that can be checked in practice.

Theorem 55 *The infinite matrix* $\mathbf{T} = (\downarrow 2)2\mathbf{H}\overline{\mathbf{H}}^{\rm tr} = \mathbf{M}\overline{\mathbf{H}}^{\rm tr}$ *and its finite submatrix* $\mathbf{T}_{2N-1}$ *always have* $\lambda = 1$ *as an eigenvalue. The cascade iteration* $\mathbf{a}^{(i+1)} = \mathbf{T}\mathbf{a}^{(i)}$ *converges in* ℓ^2 *to the eigenvector* $\mathbf{a} = \mathbf{T}\mathbf{a}$ *if and only if the following condition is satisfied:*

- *All of the eigenvalues* λ *of* $\mathbf{T}_{2N-1}$ *satisfy* $|\lambda| < 1$ *except for the simple eigenvalue* $\lambda = 1$.

PROOF: let λ_j be the $2N-1$ eigenvalues of $\mathbf{T}_{2N-1}$, including multiplicities. Then there is a basis for the space of $2N-1$-tuples with respect to which $\mathbf{T}_{2N-1}$ takes the *Jordan canonical form*

$$\tilde{\mathbf{T}}_{2N-1} = \begin{pmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & & \\ & & \lambda_p & & & & \\ & & & A_{p+1} & & & \\ & & & & A_{p+2} & & \\ & & & & & \ddots & \\ & & & & & & A_{p+q} \end{pmatrix}$$

where the Jordan blocks look like

$$A_s = \begin{pmatrix} \lambda_s & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_s & 1 & \cdots & 0 & 0 \\ \cdots & & & & & \cdots \\ 0 & 0 & 0 & \cdots & \lambda_s & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_s \end{pmatrix}.$$

If the eigenvectors of $\mathbf{T}_{2N-1}$ form a basis, for example if there were $2N-1$ distinct eigenvalues, then with respect to this basis $\tilde{\mathbf{T}}_{2N-1}$ would be diagonal and there would be no Jordan blocks. In general, however, there may not be enough eigenvectors to form a basis and the more general Jordan form will hold, with Jordan blocks. Now suppose we perform the cascade recursion n times. Then the action of the iteration on the base space will be

$$\tilde{\mathbf{T}}^n_{2N-1} = \begin{pmatrix} \lambda_1^n & & & & & & \\ & \ddots & & & & & \\ & & \lambda_p^n & & & & \\ & & & A^n_{p+1} & & & \\ & & & & A^n_{p+2} & & \\ & & & & & \ddots & \\ & & & & & & A^n_{p+q} \end{pmatrix}$$

where

$$A^n_s = \begin{pmatrix} \lambda_s^n & \binom{n}{1}\lambda_s^{n-1} & \binom{n}{2}\lambda_s^{n-2} & \cdots & \binom{n}{m_s-2}\lambda_s^{n-m_s+2} & \binom{n}{m_s-1}\lambda_s^{n-m_s+1} \\ 0 & \lambda_s^n & & \cdots & \binom{n}{m_s-3}\lambda_s^{n-m_s+3} & \binom{n}{m_s-2}\lambda_s^{n-m_s+2} \\ \cdots & & & & & \cdots \\ 0 & 0 & 0 & \cdots & \lambda_s^n & \binom{n}{1}\lambda_s^{n-1} \\ 0 & 0 & 0 & \cdots & 0 & \lambda_s^n \end{pmatrix}.$$

and A_s is an $m_s \times m_s$ matrix and m_s is the multiplicity of the eigenvalue λ_s. If there is an eigenvalue with $|\lambda_j| > 1$ then the corresponding terms in the power matrix will blow up and the cascade algorithm will fail to converge. (Of course if the original input vector has zero components corresponding to the basis vectors with these eigenvalues and the computation is done with perfect accuracy, one might have convergence. However, the slightest deviation, such as due to roundoff error, would introduce a component that would blow up after repeated iteration. Thus in practice the algorithm would diverge.The same remarks apply to Theorem 47 and Corollary 13. With perfect accuracy and filter coefficients that satisfy double-shift orthogonality, one can maintain orthogonality of the shifted scaling functions at each pass of the cascade algorithm if orthogonality holds for the initial step. However, if the algorithm diverges, this theoretical result is of no practical importance. Roundoff error would lead to meaningless results in successive iterations.)

Similarly, if there is a Jordan block corresponding to an eigenvalue $|\lambda_j| = 1$ then the algorithm will diverge. If there is no such Jordan block, but there is more than one eigenvalue with $|\lambda_j| = 1$ then there may be convergence, but it won't be unique and will differ each time the algorithm is applied. If, however, all eigenvalues satisfy $|\lambda_j| < 1$ except for the single eigenvalue $\lambda_1 = 1$, then in the limit as $n \to \infty$ we have

$$\lim_{n\to\infty} \tilde{\mathbf{T}}^n_{2N-1} = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}$$

and there is convergence to a unique limit. Q.E.D.

In the frequency domain the action of the $\mathbf{T}$ operator is

$$\mathbf{T}X(2\omega) = |H(\omega)|^2 X(\omega) + |H(\omega+\pi)|^2 X(\omega+\pi). \tag{8.7}$$

Here $X(\omega) = \sum_{n=-N+1}^{N-1} \mathbf{x}(n)e^{-in\omega}$ and $\mathbf{x}(n)$ is a $2N-1$-tuple. In the z-domain this is

$$\mathbf{T}X(z^2) = H(z)\overline{H}(z^{-1})X(z) + H(-z)\overline{H}(-z^{-1})X(-z). \tag{8.8}$$

where $H(z) = \sum_{k=0}^{N} \mathbf{h}(k)z^{-k}$ and $\overline{H}(z^{-1}) = \sum_{k=0}^{N} \overline{\mathbf{h}}(k)z^{k}$. The $\mathbf{x} \neq 0$ is an eigenvector of $\mathbf{T}$ with eigenvalue λ if and only if $\mathbf{T}\mathbf{x} = \lambda\mathbf{x}$, i.e.,

$$\lambda X(z^2) = H(z)\overline{H}(z^{-1})X(z) + H(-z)\overline{H}(-z^{-1})X(-z). \tag{8.9}$$

We can gain some additional insight into the behavior of the eigenvalues of $\mathbf{T}$ through examining it in the z-domain. Of particular interest is the effect on the eigenvalues of $p > 1$ zeros at $z = -1$ for the low pass filter $\mathbf{H}$. We can write $H(z) = (\frac{1+z^{-1}}{2})^{p-1}H_0(z)$ where $H_0(z)$ is the z-transform of the low pass filter $\mathbf{H}_0$ with a single zero at $z = -1$. In general, $\mathbf{H}_0$ won't satisfy the double-shift orthonormality condition, but we will still have $H_0(1) = 1$ and $H_0(-1) = 0$. This means that the column sums of $\mathbf{T}_0$ are equal to 1 so that in the time domain $\mathbf{T}_0$ admits the left-hand eigenvector $\mathbf{e_0} = (1, 1, \cdots, 1)$ with eigenvalue 1. Thus $\mathbf{T}_0$ also has some right-hand eigenvector with eigenvalue 1. Here, $\mathbf{T}_0$ is acting on a $2N_0 - 1$ dimensional space, where $N_0 = N - 2(p-1)$

Our strategy will be to start with $H_0(z)$ and then successively multiply it by the $p-1$ terms $(\frac{1+z^{-1}}{2})$, one-at-a-time, until we reach $H(z)$. At each stage we will use equation (8.9) to track the behavior of the eigenvalues and eigenvectors. Each time we multiply by the factor we will add 2 dimensions to the space on which we are acting. Thus there will be 2 additional eigenvalues at each recursion. Suppose we have reached stage $H_s(z) = (\frac{1+z^{-1}}{2})^s H_0(z)$ in this process, with $0 \le s < p-1$. Let $\mathbf{x}_s$ be an eigenvector of the corresponding operator $\mathbf{T}_s$ with eigenvalue λ_s. In the z-domain we have

$$\lambda_s X_s(z^2) = H_s(z)\overline{H}_s(z^{-1})X_s(z) + H_s(-z)\overline{H}_s(-z^{-1})X_s(-z). \tag{8.10}$$

Now let $H_{s+1}(z) = (\frac{1+z^{-1}}{2})H_s(z)$, $X_{s+1}(z) = (1-z^{-1})(1-z)X_s(z)$. Then, since

$$(\frac{1+z^{-1}}{2})(\frac{1+z}{2})(1-z^{-1})(1-z) = \frac{1}{4}(1-z^{-2})(1-z^2)$$

the eigenvalue equation transforms to

$$\frac{1}{4}\lambda_s X_{s+1}(z^2) = H_{s+1}(z)\overline{H}_{s+1}(z^{-1})X_{s+1}(z) + H_{s+1}(-z)\overline{H}_{s+1}(-z^{-1})X_{s+1}(-z).$$

Thus, each eigenvalue λ_s of $\mathbf{T}_s$ transforms to an eigenvalue $\lambda_s/4$ of $\mathbf{T}_{s+1}$. In the time domain, the new eigenvectors are linear combinations of shifts of the old ones. There are still 2 new eigenvalues and their associated eigenvectors to be accounted for. One of these is the eigenvalue 1 associated with the left-hand eigenvector $\mathbf{e_0} = (1, 1, \cdots, 1)$. (The right-hand eigenvector is the all important $\mathbf{a}$.) To find the last eigenvalue and eigenvector, we consider an intermediate step between H_s and H_{s+1}.

Let

$$K_{s+1/2}(z) = (\frac{1+z^{-1}}{2})H_s(z)\overline{H}_s(z^{-1}), \quad X_{s+1/2}(z) = (1-z^{-1})X_s(z).$$

Then, since

$$(\frac{1+z^{-1}}{2})(1-z^{-1}) = \frac{1}{2}(1-z^{-2})$$

the eigenvalue equation transforms to

$$\frac{1}{2}\lambda_s X_{s+1/2}(z^2) = K_{s+1/2}(z)X_{s+1/2}(z) + K_{s+1/2}(-z)X_{s+1/2}(-z). \tag{8.11}$$

This equation doesn't have the same form as the original equation, but in the time domain it corresponds to

$$(\downarrow 2)\mathbf{K}_{s+1/2}\mathbf{x}_{s+1/2} = \frac{1}{2}\lambda \mathbf{x}_{s+1/2}.$$

The eigenvectors of $\mathbf{H}_s$ transform to eigenvectors of $(\downarrow 2)\mathbf{K}_{s+1/2}$ with halved eigenvalues. Since $K_{s+1/2}(1) = 1, K_{s+1/2}(-1) = 0$. the columns of $(\downarrow 2)\mathbf{K}_{s+1/2}$ sum to 1, and $(\downarrow 2)\mathbf{K}_{s+1/2}$ has a left-hand eigenvector $\mathbf{e_0} = (1, 1, \cdots, 1)$ with eigenvalue 1. Thus it also has a new right-hand eigenvector with eigenvalue 1. Now we repeat this process for $(\frac{1+z}{2})K_{s+1/2}(z)$, which gets us back to the eigenvalue problem for $\mathbf{H}_{s+1}$. Since existing eigenvalues are halved by this process, the new eigenvalue 1 for $(\downarrow 2)\mathbf{K}_{s+1/2}$ becomes the eigenvalue $\frac{1}{2}$ for $\mathbf{T}_s$.

NOTE: One might think that there could be a right-hand eigenvector of $(\downarrow 2)\mathbf{K}_{s+1/2}$ with eigenvalue 2 that transforms to the right-hand eigenvector with eigenvalue 1 and also that the new eigenvector or generalized eigenvector that is added to the space might then be associated with some eigenvalue $\lambda \neq 1$. However, this cannot happen; the new vector added is always associated to eigenvalue 1. First observe that the subspace $(1 - x^{-1})X_s(z)$ is invariant under the action of $(\downarrow 2)\mathbf{K}_{s+1/2}$. Thus all of these functions satisfy $X(1) = 0$. The spectral resolution of $(\downarrow 2)\mathbf{K}_{s+1/2}$ into Jordan form must include vectors $X(z)$ such that $X(1) \neq 0$. If $X(z)$ is an eigenvector corresponding to eigenvalue $\lambda \neq 1$ then the obvious modification of the eigenvalue equation (8.11) when restricted to $z = 1$ leads to the condition $\lambda X(1) = X(1)$. Hence $X(1) = 0$. It follows that vectors $X(z)$ such that $X(1) \neq 0$ can only be associated with the generalized eigenspace with eigenvalue $\lambda = 1$. Thus at each step in the process above we are always adding a new to the space and this vector corresponds to eigenvalue 1.

Theorem 56 *If $H(z)$ has a zero of order p at $z = -1$ then $\mathbf{T}$ has eigenvalues $1, \frac{1}{2}, \cdots (\frac{1}{2})^{2p-1}$.*

Theorem 57 *Assume that $\phi(t) \in L^2[-\infty, \infty]$. Then the cascade sequence $\phi^{(i)}(t)$ converges in L^2 to $\phi(t)$ if and only if the convergence criteria of Theorem 55 hold:*

- *All of the eigenvalues λ of $\mathbf{T}_{2N-1}$ satisfy $|\lambda| < 1$ except for the simple eigenvalue $\lambda = 1$.*

PROOF: Assume that the convergence criteria of Theorem 55 hold. We want to show that

$$||\phi^{(i)} - \phi||^2 = ||\phi^{(i)}||^2 - (\phi^{(i)}, \phi) - (\phi, \phi^{(i)}) + ||\phi||^2$$

$$= \mathbf{a}^{(i)}(0) - \overline{\mathbf{b}}^{(i)}(0) - \mathbf{b}^{(i)}(0) + \mathbf{a}(0) \to 0$$

as $i \to \infty$, see (8.3), (8.6). Here

$$\mathbf{a}^{(i)}(k) = \int_{-\infty}^{\infty} \phi^{(i)}(t)\overline{\phi}^{(i)}(t-k)dt, \qquad \mathbf{b}^{(i)}(k) = \int_{-\infty}^{\infty} \phi(t)\overline{\phi}^{(i)}(t-k)dt.$$

With the conditions on $\mathbf{T}$ we know that each of the vector sequences $\mathbf{a}^{(i)}$, $\mathbf{b}^{(i)}$ will converge to a multiple of the vector $\mathbf{a}$ as $i \to \infty$. Since $\mathbf{a}(k) = \delta_{0k}$ we have $\lim_{i\to\infty} \mathbf{a}^{(i)}(k) = \mu\delta_{0k}$ and $\lim_{i\to\infty} \mathbf{b}^{(i)}(k) = \nu\delta_{0k}$. Now at each stage of the recursion we have $\sum_k \mathbf{a}^{(i)}(k) = \sum_k \mathbf{b}^{(i)}(k) = 1$, so $\mu = \nu = 1$. Thus as $i \to \infty$ we have

$$\mathbf{a}^{(i)}(0) - \overline{\mathbf{b}}^{(i)}(0) - \mathbf{b}^{(i)}(0) + \mathbf{a}(0) \to 1 - 1 - 1 + 1 = 0.$$

(Note: This argument is given for orthogonal wavelets where $\mathbf{a}(k) = \delta_{0k}$. However, a modification of the argument works for biorthogonal wavelets as well. Indeed the normalization condition $\sum_k \mathbf{a}^{(i)}(k) = \sum_k \mathbf{b}^{(i)}(k) = 1$ holds also in the biorthogonal case.) Conversely, this sequence can only converge to zero if the iterates of $\mathbf{T}$ converge uniquely, hence only if the convergence criteria of Theorem 55 are satisfied. Q.E.D.

Theorem 58 *If the convergence criteria of Theorem 55 hold, then the $\phi^{(i)}(t)$ are a Cauchy sequence in L^2, converging to $\phi(t)$.*

PROOF: We need to show only that the $\phi^{(i)}(t)$ are a Cauchy sequence in L^2. Indeed, since L^2 is complete, the sequence must then converge to some $\phi \in L^2$. We have

$$||\phi^{(m)} - \phi^{(i)}||^2 = ||\phi^{(m)}||^2 - (\phi^{(m)}, \phi^{(i)}) - (\phi^{(i)}, \phi^{(m)}) + ||\phi^{(i)}||^2.$$

From the proof of the preceding theorem we know that $\lim_{i\to\infty} ||\phi^{(i)}||^2 = 1$. Set $m = i + j$ for fixed $j > 0$ and define the vector

$$\mathbf{c}_j^{(i)}(k) = (\phi_{00}^{(i+j)}, \phi_{0k}^{(i)})$$

$$= \int_{-\infty}^{\infty} \phi^{(i+j)}(t)\overline{\phi}^{(i)}(t-k)dt,$$

i.e., the vector of inner products at stage i. A straight-forward computation yields the recursion

$$\mathbf{c}_j^{(i+1)} = \mathbf{T}\mathbf{c}_j^{(i)}.$$

Since $\sum_k \mathbf{c}_j^{(i+1)}(k) = 1$ at each stage i in the recursion, it follows that $\lim_{i\to\infty} \mathbf{c}_j^{(i)}(k) = \mathbf{a}(k)$ for each j. The initial vectors for these recursions are $\mathbf{c}_j^{(0)}(k) = \int_{-\infty}^{\infty} \phi^{(j)}(t)\overline{\phi}^{(0)}(t-k)dt$. We have $\sum_k \mathbf{c}_j^{(0)}(k) = 1$ so $\mathbf{c}_j^{(i)}(k) \to \mathbf{a}(k)$ as $i \to \infty$. Furthermore, by the Schwarz inequality $|\mathbf{c}_j^{(0)}(k)| \leq ||\phi^{(j)}|| \cdot ||\phi^{(0)}|| = 1$, so the components are uniformly bounded. Thus $\mathbf{T}^i\mathbf{c}_j^{(0)}(0) \to \mathbf{a}(0) = 1$ as $i \to \infty$, uniformly in j. It follows that

$$||\phi^{(i+j)} - \phi^{(i)}||^2 = ||\phi^{(i+j)}||^2 - (\phi^{(i+j)}, \phi^{(i)}) - (\phi^{(i)}, \phi^{(i+j)}) + ||\phi^{(i)}||^2 \to 0$$

as $i \to \infty$, uniformly in j. Q.E.D.

We continue our examination of the eigenvalues of $\mathbf{T}$ particularly in cases related to Daubechies wavelets. We have observed that in the frequency domain an eigenfunction X corresponding to eigenvalue λ of the $\mathbf{T}$ operator is characterized by the equation

$$\lambda X(\omega) = |H(\frac{\omega}{2})|^2 X(\frac{\omega}{2}) + |H(\frac{\omega}{2}+\pi)|^2 X(\frac{\omega}{2}+\pi), \tag{8.12}$$

where $X(\omega) = \sum_{n=-N}^{N} \mathbf{x}(n)e^{-in\omega}$ and $\mathbf{x}(n)$ is a $2N-1$-tuple. We normalize X by requiring that $||X|| = 1$.

Theorem 59 *If $H(\omega)$ satisfies the conditions*

- $$|H(\omega)|^2 + |H(\omega+\pi)|^2 \equiv 1,$$
- $$H(0) = 1, \qquad H(\pi) = 0,$$
- $$|H(\omega)| \neq 0 \text{ for } -\pi/2 < \omega < \pi/2,$$

then $\mathbf{T}$ has a simple eigenvalue 1 and all other eigenvalues satisfy $|\lambda| < 1$.

PROOF: The key to the proof is the observation that for any fixed ω with $0 < |\omega| < \pi$ we have $\lambda X(\omega) = \alpha X(\frac{\omega}{2}) + \beta X(\frac{\omega}{2}+\pi)$ where $\alpha > 0, \beta \geq 0$ and $\alpha + \beta = 1$. Thus $\lambda X(\omega)$ is a weighted average of $X(\frac{\omega}{2})$ and $X(\frac{\omega}{2}+\pi)$.

- There are no eigenvalues with $|\lambda| > 1$. For, suppose X were a normalized eigenvector corresponding to λ. Also suppose $|X(\omega)|$ takes on its maximum value at ω_0, such that $0 < \omega_0 < 2\pi$. Then $|\lambda| \cdot |X(\omega_0)| \leq \alpha|X(\frac{\omega_0}{2})| + \beta|X(\frac{\omega_0}{2} + \pi)|$, so, say, $|\lambda| \cdot |X(\omega_0)| \leq |X(\frac{\omega_0}{2})|$. Since $|\lambda| > 1$ this is impossible unless $|X(\omega_0)| = 0$. On the other hand, setting $\omega = 0$ in the eigenvalue equation we find $\lambda X(0) = X(0)$, so $X(0) = 0$. Hence $X(\omega) \equiv 0$ and λ is not an eigenvalue.

- There are no eigenvalues with $|\lambda| = 1$ but $\lambda \neq 1$. For, suppose X was a normalized eigenvector corresponding to λ. Also suppose $|X(\omega)|$ takes on its maximum value at ω_0, such that $0 < \omega_0 < 2\pi$. Then $|\lambda| \cdot |X(\omega_0)| \leq \alpha|X(\frac{\omega_0}{2})| + \beta|X(\frac{\omega_0}{2} + \pi)|$, so $|X(\omega_0)| = |X(\frac{\omega_0}{2})|$. (Note: This works exactly as stated for $0 < \omega_0 < \pi$. If $\pi < \omega_0 < 2\pi$ we can replace ω_0 by $\omega_0' = \omega_0 - 2\pi$ and argue as before. The same remark applies to the cases to follow.) Furthermore, $X(\omega_0) = \lambda^{-1} X(\frac{\omega_0}{2})$. Repeating this argument n times we find that $X(\frac{\omega}{2^n}) = \lambda^n X(\omega_0)$. Since $X(\omega)$ is a continuous function, the left-hand side of this expression is approaching $X(0)$ in the limit. Further, setting $\omega = 0$ in the eigenvalue equation we find $\lambda X(0) = X(0)$, so $X(0) = 0$. Thus $|X(\omega_0)| = 0$ and λ is not an eigenvalue.

- $\lambda = 1$ is an eigenvalue, with the unique (normalized) eigenvector $X(\omega) = \frac{1}{\sqrt{2\pi}}$. Indeed, for $\lambda = 1$ we can assume that $X(\omega)$ is real, since both $X(\omega)$ and $\overline{X}(\omega)$ satisfy the eigenvalue equation. Now suppose the eigenvector X takes on its maximum positive value at ω_0, such that $0 < \omega_0 < 2\pi$. Then $X(\omega_0) \leq \alpha X(\frac{\omega_0}{2}) + \beta X(\frac{\omega_0}{2} + \pi)$, so $X(\omega_0) = X(\frac{\omega_0}{2})$. Repeating this argument n times we find that $X(\frac{\omega}{2^n}) = X(\omega_0)$. Since $X(\omega)$ is a continuous function, the left-hand side of this expression is approaching $X(0)$ in the limit. Thus $X(\omega_0) = X(0)$. Now repeat the same argument under the supposition that the eigenvector X takes on its minimum value at ω_1, such that $0 < \omega_1 < 2\pi$. We again find that $X(\omega_1) = X(0)$. Thus, $X(\omega)$ is a constant function. We already know that this constant function is indeed an eigenvector with eigenvalue 1.

- There is no nontrivial Jordan block corresponding to the eigenvalue 1. Denote the normalized eigenvector for eigenvalue 1, as computed above, by $X_1(\omega) = \frac{1}{\sqrt{2\pi}}$. If such a block existed there would be a function $X(\omega)$, not normalized in general, such that $\mathbf{T}X(\omega) = X(\omega) + X_1(\omega)$, i.e.,

$$X(\omega) + \frac{1}{\sqrt{2\pi}} = |H(\frac{\omega}{2})|^2 X(\frac{\omega}{2}) + |H(\frac{\omega}{2} + \pi)|^2 X(\frac{\omega}{2} + \pi).$$

Now set $\omega = 0$. We find $X(0) + \frac{1}{\sqrt{2\pi}} = X(0)$ which is impossible. Thus there is no nontrivial Jordan block for $\lambda = 1$.

Q.E.D.

NOTE: It is known that the condition $|H(\omega)| \neq 0$ for $-\pi/2 < \omega < \pi/2$, can be relaxed to just hold for $-\pi/3 < \omega < \pi/3$.

From our prior work on the maxflat (Daubechies) filters we saw that $|H(\omega)|^2$ is 1 for $\omega = 0$ and decreases (strictly) monotonically to 0 for $\omega = \pi$. In particular, $|H(\omega)| \neq 0$ for $0 \leq \omega < \pi$. Since $|H(\omega)|^2 = 1 - |H(\omega+\pi)|^2$ we also have $|H(\omega)| \neq 0$ for $0 \geq \omega > -\pi$. Thus the conditions of the preceding theorem are satisfied, and the cascade algorithm converges for each maxflat system to yield the Daubechies wavelets D_N, where $N = 2p-1$ and p is the number of zeros of $H(\omega)$ at $\omega = \pi$. The scaling function is supported on the interval $[0, N]$. Polynomials of order $\leq p$ can be approximated with no error in the wavelet space V_0.

8.2 Accuracy of approximation

In this section we assume that the criteria for the eigenvalues of $\mathbf{T}$ are satisfied and that we have a multiresolution system with scaling function $\phi(t)$ supported on $[0, N]$. The related low pass filter transform function $H(\omega)$ has $p > 0$ zeros at $\omega = \pi$. We know that

$$\sum_k \mathbf{y}_{\ell k}\phi(t+k) = t^\ell, \qquad \ell = 0, 1, \cdots, p-1,$$

so polynomials in t of order $\leq p-1$ can be expressed in V_0 with no error.

Given a function $f(t)$ we will examine how well f can be approximated pointwise by wavelets in V_j, as well as approximated in the L^2 sense. We will also look at the rate of decay of the wavelet coefficients b_{jk} as $j \to \infty$. Clearly, as j grows the accuracy of approximation of $f(t)$ by wavelets in V_j grows, but so does the computational difficulty. We will not go deeply into approximation theory, but far enough so that the basic dependence of the accuracy on j and p will emerge. We will also look at the smoothness of wavelets, particularly the relationship between smoothness and p.

Let's start with pointwise convergence. Fix $j = J$ and suppose that f has p continuous derivatives in the neighborhood $|t - t_0| \leq \frac{1}{2^J}$ of t_0. Let

$$f_J(t) = \sum_k a_{Jk}\phi_{Jk}(t) = \sum_k a_{Jk}2^{J/2}\phi(2^J t - k),$$

$$a_{Jk} = (f, \phi_{Jk}) = 2^{J/2} \int_{-\infty}^{\infty} f(t)\overline{\phi}(2^J t - k)dt,$$

be the projection of f on the scaling space V_J. We want to estimate the pointwise error $|f(t) - f_J(t)|$ in the neighborhood $|t - t_0| \leq \frac{1}{2^J}$.

Recall Taylor's theorem from basic calculus.

Theorem 60 *If $f(t)$ has p continuous derivatives on an interval containing t_0 and t, then*

$$f(t) = \sum_{k=0}^{p-1} f^{(k)}(t_0)\frac{(t-t_0)^k}{k!} + R_p(t, t_0), \qquad R_p(t, t_0) = \int_{t_0}^{t} \frac{(t-x)^p}{p!} f^{(p)}(x)dx,$$

where $f^{(0)}(t) = f(t)$.

Since all polynomials of order $\leq p - 1$ can be expressed exactly in V_J we can assume that the first p terms in the Taylor expansion of f have already been canceled exactly by terms a'_{Jk} in the wavelet expansion. Thus

$$|f(t) - f_J(t)| = |R_p(t, t_0) - \sum_k a''_{Jk}\phi_{Jk}(t)|,$$

where the $a''_{Jk} = (R_p(t, t_0), \phi_{Jk})$ are the remaining coefficients in the wavelet expansion ($a_{Jk} = a'_{Jk} + a''_{Jk}$). Note that for fixed t the sum in our error expression contains only N nonzero terms at most. Indeed, the support of $\phi(t)$ is contained in $[0, N)$, so the support of $\phi_{Jk}(t)$ is contained in $[\frac{k}{2^J}, \frac{k+N}{2^J}]$. Then $\phi_{Jk}(t) = 0$ unless $k = [2^J t] - \ell$, $\ell = 0, 1, \cdots, N - 1$, where $[x]$ is the greatest integer in x. If $|f^{(p)}|$ has upper bound M_p in the interval $|t - t_0| \leq \frac{1}{2^J}$ then

$$|R_p(t, t_0)| \leq \frac{M_p}{2^{J(p+1)}(p+1)!}$$

and we can derive similar upper bounds for the other N terms that contribute to the sum, to obtain

$$|f(t) - f_J(t)| \leq \frac{CM_p}{2^{J(p+1)}} \tag{8.13}$$

where C is a constant, independent of f and J. If f has only $\ell < p$ continuous derivatives in the interval $|t - t_0| \leq \frac{1}{2^J}$ then the p in estimate (8.13) is replaced by ℓ:

$$|f(t) - f_J(t)| \leq \frac{CM_\ell}{2^{J(\ell+1)}}.$$

Note that this is a *local* estimate; it depends on the smoothness of f in the interval $|t-t_0| \le \frac{1}{2^J}$. Thus once the wavelets choice is fixed, the local rate of convergence can vary dramatically, depending only on the local behavior of f. This is different from Fourier series or Fourier integrals where a discontinuity of a function at one point can slow the rate of convergence at all points. Note also the dramatic improvement of convergence rate due to the p zeros of $H(\omega)$ at $\omega = \pi$.

It is interesting to note that we can investigate the pointwise convergence in a manner very similar to the approach to Fourier series and Fourier integral pointwise convergence in the earlier sections of these notes. Since $\sum_k \phi(t+k) = 1$ and $\int \phi(t)dt = 1$ we can write (for $0 \le t \le \frac{1}{2^J}$.)

$$|f(t) - f_J(t)| = |f(t) - \sum_k 2^J \int_{-\infty}^{\infty} f(x)\overline{\phi}(2^J x - k)dx\phi(2^J t - k)|$$

$$= |\sum_k \left(f(t) - 2^J \int_{-\infty}^{\infty} f(x)\overline{\phi}(2^J x - k)dx \right) \phi(2^J t - k)|$$

$$\le \sup|\phi| \sum_{k=-N+1}^{0} \left| \int_{-\infty}^{\infty} \left[f(t) - f(\frac{u+k}{2^J}) \right] \overline{\phi}(u)du \right|. \qquad (8.14)$$

Now we can make various assumptions concerning the smoothness of f to get an upper bound for the right-hand side of (8.14). We are not taking any advantage of special features of the wavelets employed. Here we assume that f is continuous everywhere and has finite support. Then, since f is uniformly continuous on its domain, it is easy to see that the following function exists for every $h \ge 0$:

$$\omega(h) = \sup_{|t-t'|\le h,\ t,t' \text{ real}} |f(t) - f(t')|. \qquad (8.15)$$

Clearly, $\omega(h) \to 0$ as $h \to 0$. We have the bound

$$|f(t) - f_J(t)| \le \sup|\phi| \cdot N\omega(\frac{N}{2^J}) \int_0^N |\phi(u)|du.$$

If we repeat this computation for $t \in [\frac{k}{2^J}, \frac{k+1}{2^J}]$ we get the same upper bound. Thus this bound is uniform for all t, and shows that $f_J(t) \to f(t)$ uniformly as $J \to \infty$.

Now we turn to the estimation of the wavelet expansion coefficients

$$b_{jk} = (f, w_{jk}) = 2^{j/2} \int_{-\infty}^{\infty} f(t)\overline{w}(2^j t - k)dt \qquad (8.16)$$

where $w(t)$ is the mother wavelet. We could use Taylor's theorem for f here too, but I will present an alternate approach. Since $w(t)$ is orthogonal to all integer translates of the scaling function $\phi(t)$ and since all polynomials of order $\leq p-1$ can be expressed in V_0, we have

$$\int_{-\infty}^{\infty} t^\ell w(t)dt = 0, \qquad \ell = 0, 1, \cdots, p-1. \tag{8.17}$$

Thus the first p moments of w vanish. We will investigate some of the consequences of the vanishing moments. Consider the functions

$$\begin{aligned} I_1(t) &= \int_{-\infty}^{t} w(\tau_0)d\tau_0 \\ I_2(t) &= \int_{-\infty}^{t} I_1(\tau_1)d\tau_1 = \int_{-\infty}^{t} d\tau_1 \int_{-\infty}^{\tau_1} w(\tau_0)d\tau_0 \\ I_m(t) &= \int_{-\infty}^{t} I_{m-1}(\tau_{m-1})d\tau_{m-1}, \quad m = 1, 2, \cdots p. \end{aligned} \tag{8.18}$$

From equation (8.17) with $\ell = 0$ it follows that $I_1(t)$ has support contained in $[0, N)$. Indeed we can always arrange matters such that the support of $w(t)$ is contained in $[0, N)$. Thus $\int_0^N t^\ell w(t)dt = 0$. Integrating by parts the integral in equation (8.17) with $\ell = 1$, it follows that $I_2(t)$ has support contained in $[0, N)$. We can continue integrating by parts in this series of equations to show, eventually, that $I_p(t)$ has support contained in $[0, N)$. (This is as far as we can go, however.)

Now integrating by parts p times we find

$$b_{jk} = 2^{j/2} \int_{-\infty}^{\infty} f(t)\overline{w}(2^j t - k)dt = 2^{-j/2} \int_{-\infty}^{\infty} f(\frac{u+k}{2^j})\overline{w}(u)du =$$

$$2^{-j/2}(-1)^p \int_{-\infty}^{\infty} \frac{d^p}{du^p} f(\frac{u+k}{2^j})\overline{I}_p(u)du = 2^{j/2-pj}(-1)^p \int_{-\infty}^{\infty} f^{(p)}(\frac{u+k}{2^j})\overline{I}_p(u)du$$

$$= 2^{j/2-pj}(-1)^p \int_{-\infty}^{\infty} f^{(p)}(t)\overline{I}_p(2^j t - k)dt$$

$$= 2^{j/2-pj}(-1)^p \int_{\frac{k}{2^j}}^{\frac{N+k}{2^j}} f^{(p)}(t)\overline{I}_p(2^j t - k)dt.$$

If $|f^{(p)}(t)|$ has a uniform upper bound M_p then we have the estimate

$$|b_{jk}| \leq \frac{CM_p}{2^{(p+1/2)j}} \tag{8.19}$$

where C is a constant, independent of f, j, k. If, moreover, $f(t)$ has bounded support, say within the interval $(0, K)$ then we can assume that $b_{jk} = 0$ unless

$0 \leq k \leq 2^j K$. We already know that the wavelet basis is complete in $L^2[-\infty, \infty]$. Let's consider the decomposition

$$L^2[-\infty, \infty] = V_J \oplus \sum_{j=J}^{\infty} W_j.$$

We want to estimate the L^2 error $||f - f_J||$ where f_J is the projection of f on V_J. From the Plancherel equality and our assumption that the support of f is bounded, this is

$$||f - f_J||^2 = \sum_{j=J}^{\infty} \sum_{k=0}^{2^j K} |b_{jk}|^2 < \frac{2C^2 M_p^2 K}{2^{2pJ}}. \tag{8.20}$$

Again, if f has $1 \leq \ell < p$ continuous derivatives then the p in (8.20) is replaced by ℓ.

Much more general and sophisticated estimates than these are known, but these provide a good guide to the convergence rate dependence on j, p and the smoothness of f.

Next we consider the estimation of the scaling function expansion coefficients

$$a_{jk} = (f, \phi_{jk}) = 2^{j/2} \int_{-\infty}^{\infty} f(t)\overline{\phi}(2^j t - k)dt \tag{8.21}$$

In order to start the FWT recursion for a function f, particularly a continuous function, it is very common for people to choose a large $j = J$ and then use function samples to approximate the coefficients: $a_{Jk} \sim 2^{-J/2} f(\frac{k}{2^J})$. This may not be a good policy. Let's look more closely. Since the support of $\phi(t)$ is contained in $[0, N)$ the integral for a_{jk} becomes

$$a_{jk} = 2^{j/2} \int_{\frac{k}{2^j}}^{\frac{N+k}{2^j}} f(t)\overline{\phi}(2^j t - k)dt = 2^{-j/2} \int_0^N f(\frac{u+k}{2^j})\overline{\phi}(u)du.$$

The approximation above is the replacement

$$\int_0^N f(\frac{u+k}{2^j})\overline{\phi}(u)du \sim f(\frac{k}{2^j}) \int_0^N \overline{\phi}(u)du = f(\frac{k}{2^j}).$$

If j is large and f is continuous this using of samples of f isn't a bad estimate. If f is discontinuous or only defined at dyadic values, the sampling could be wildly inaccurate. Note that if you start the FWT recursion at $j = J$ then

$$f_J(t) = \sum_k a_{Jk}\phi_{Jk}(t),$$

so it would be highly desirable for the wavelet expansion to correctly fit the sample values at the points $\frac{\ell}{2^J}$:

$$f(\frac{\ell}{2^J}) = f_J(\frac{\ell}{2^J}) = \sum_k a_{Jk}\phi_{Jk}(\frac{\ell}{2^J}). \tag{8.22}$$

However, if you use the sample values $f(\frac{k}{2^J})$ for a_{Jk} then in general $f_J(\frac{\ell}{2^J})$ will not reproduce the sample values! Strang and Nyuyen recommend prefiltering the samples to ensure that (8.22) holds. For Daubechies wavelets, this amounts to replacing the integral $a_{Jk} = (f, \phi_{Jk})$ by the sum $a^c_{Jk} = \sum_\ell f(\frac{\ell}{2^J})\phi_{Jk}(\frac{\ell}{2^J})$. These "corrected" wavelet coefficients a^c_{Jk} will reproduce the sample values. There is no unique, final answer as to how to determine the initial wavelet coefficients. The issue deserves some thought, rather than a mindless use of sample values.

8.3 Smoothness of scaling functions and wavelets

Our last major issue in the construction of scaling functions and wavelets via the cascade algorithm is their smoothness. So far we have shown that the Daubechies scaling functions are in $L^2[-\infty, \infty]$. We will use the method of Theorem 56 to examine this. The basic result is this: The matrix $\mathbf{T}$ has eigenvalues $1, \frac{1}{2}, \frac{1}{4}, \cdots, \frac{1}{2^{2p-1}}$ associated with the zeros of $H(\omega)$ at $\omega = \pi$. If all other eigenvalues λ of $\mathbf{T}$ satisfy $|\lambda| < \frac{1}{4^s}$ then $\phi(t)$ and $w(t)$ have s derivatives. We will show this for integer s. It is also true for fractional derivatives s, although we shall not pursue this.

Recall that in the proof of Theorem 56 we studied the effect on $\mathbf{T}$ of multiplying $H(z)$ by factors $\frac{1+z^{-1}}{2}$, each of which adds a zero at $z = -1$. We wrote $H(z) = (\frac{1+z^{-1}}{2})^{p-1}H_0(z)$ where $H_0(z)$ is the z-transform of the low pass filter $\mathbf{H}_0$ with a single zero at $z = -1$. Our strategy was to start with $H_0(z)$ and then successively multiply it by the $p-1$ terms $(\frac{1+z^{-1}}{2})$, one-at-a-time, until we reached $H(z)$. At each stage, every eigenvalue λ_i of the preceding matrix $\mathbf{T}_i$ transformed to an eigenvalue $\lambda_i/4$ of $\mathbf{T}_{i+1}$. There were two new eigenvalues added (1 and $1/2$), associated with the new zero of $H(\omega)$.

In going from stage i to stage $i+1$ the infinite product formula for the scaling function

$$\hat{\phi}_{(i)}(\omega) = \Pi_{j=1}^{\infty} H_i(\frac{\omega}{2^j}), \tag{8.23}$$

changes to

$$\hat{\phi}_{(i+1)}(\omega) = \Pi_{j=1}^{\infty} H_{i+1}(\frac{\omega}{2^j}) = \Pi_{j=1}^{\infty}\left(\frac{1}{2} + \frac{1}{2}e^{-i\omega/2^j}\right)\Pi_{j=1}^{\infty} H_i(\frac{\omega}{2^j})$$

$$= \left(\frac{1 - e^{-i\omega}}{i\omega} \right) \hat{\phi}_{(i)}(\omega). \tag{8.24}$$

The new factor is the Fourier transform of the box function. Now suppose that $\mathbf{T}_i$ satisfies the condition that all of its eigenvalues λ_i are < 1 in absolute value, except for the simple eigenvalue 1. Then $\hat{\phi}_{(i)}(\omega) \in L^2$, so $\int_{-\infty}^{\infty} |\omega|^2 |\hat{\phi}_{(i+1)}(\omega)|^2 d\omega < \infty$. Now

$$\phi_{(i+1)}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\phi}_{(i+1)}(\omega) e^{i\omega t} d\omega$$

and the above inequality allows us to differentiate with respect to t under the integral sign on the right-hand side:

$$\phi'_{(i+1)}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} i\omega \hat{\phi}_{(i+1)}(\omega) e^{i\omega t} d\omega.$$

The derivative not only exists, but by the Plancherel theorem, it is square integrable. Another way to see this (modulo a few measure theoretic details) is in the time domain. There the scaling function $\phi_{(i+1)}(t)$ is the convolution of $\phi_{(i)}(t)$ and the box function:

$$\phi_{(i+1)}(t) = \int_0^1 \phi_{(i)}(t - x) dx = \int_{t-1}^{t} \phi_{(i)}(u) du.$$

Hence $\phi'_{(i+1)}(t) = \phi_{(i)}(t) - \phi_{(i)}(t-1)$. (Note: Since $\phi_{(i)}(t) \in L^2$ it is locally integrable.) Thus $\phi_{(i+1)}(t)$ is differentiable and it has one more derivative than $\phi_{(i+1)}(t)$. Thus once we have (the non-special) eigenvalues of $\mathbf{T}_i$ less than one in absolute value, each succeeding zero of H adds a new derivative to the scaling function.

Theorem 61 *If all eigenvalues λ of $\mathbf{T}$ satisfy $|\lambda| < \frac{1}{4^s}$ (except for the special eigenvalues $1, \frac{1}{2}, \frac{1}{4}, \cdots, \frac{1}{2^{2p-1}}$, each of multiplicity one) then $\phi(t)$ and $w(t)$ have s derivatives.*

Corollary 18 *The convolution $\phi_{i+1}(t)$ has $k+1$ derivatives if and only if $\phi_k(t)$ has k derivatives.*

PROOF: From the proof of the theorem, if $\phi_i(t)$ has k derivatives, then $\phi_{i+1}(t)$ has one more derivative. Conversely, suppose $\phi_{i+1}(t)$ has $k+1$ derivatives. Then $\phi'_{i+1}(t)$ has k derivatives, and from (8.3) we have

$$\phi'_{i+1}(t) = \phi_i(t) - \phi_i(t-1).$$

Thus, $\phi_i(t) - \phi_i(t-1)$ has k derivatives. Now $\phi_i(t)$ corresponds to the FIR filter H_i so $\phi_i(t)$ has support in some bounded interval $[0, M)$. Note that the function

$$\phi_i(t) - \phi_i(t-M) = \sum_{j=1}^{M} \left(\phi_i(t) - \phi_i(t-j)\right).$$

must have k derivatives, since each term on the right-hand side has k derivatives. However, the support of $\phi_i(t)$ is disjoint from the support of $\phi_i(t-M)$, so $\phi_i(t)$ itself must have k derivatives. Q.E.D.

Corollary 19 *If $\phi(t)$ has s derivatives in L^2 then $s < p$. Thus the maximum possible smoothness is $p-1$.*

EXAMPLES:

- Daubechies D_4. Here $N = 3, p = 2$, so the T matrix is 5×5 ($2N-1 = 5$), or T_5. Since $p = 2$ we know four roots of T_5: $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$. We can use Matlab to find the remaining root. It is $\lambda = \frac{1}{4}$. This just misses permitting the scaling function to be differentiable; we would need $|\lambda| < \frac{1}{4}$ for that. Indeed by plotting the D_4 scaling function using the Matlab toolbox, or looking up to graph of this function in your text, you can see that there are definite corners in the graph. Even so, it is less smooth than it appears. It can be shown that, in the frequency domain, $\int_{-\infty}^{\infty} |\omega|^{2s} |\hat{\phi}(\omega)|^2 d\omega < \infty$ for $0 \leq s < 1$ (but not for $s = 1$). This implies continuity of $\phi(t)$, but not quite differentiability.

- General Daubechies D_M. Here $M = N + 1 = 2p$. By determining the largest eigenvalue in absolute value other than the known $2p$ eigenvalues $1, \frac{1}{2}, \cdots, \frac{1}{2^{2p-1}}$ one can compute the number of derivatives s admitted by the scaling functions. The results for the smallest values of p are as follows. For $p = 1, 2$ we have $s = 0$. For $p = 3, 4$ we have $s = 1$. For $p = 5, 6, 7, 8$ we have $s = 2$. For $p = 9, 10$ we have $s = 3$. Asymptotically s grows as $0.2075p + \text{ constant}$.

We can derive additional smoothness results by considering more carefully the pointwise convergence of the cascade algorithm in the frequency domain:

$$\hat{\phi}(\omega) = \Pi_{j=1}^{\infty} H(\frac{\omega}{2^j}).$$

Recall that we only had the crude upper bound $|\hat{\phi}(\omega)| \leq e^{C_0|\omega|}$ where $|H'(\omega)| \leq C_0$ and $H(\omega) = H(0) + \int_0^\omega H'(s)ds$, so

$$|H(\omega)| \leq 1 + C_0|\omega| \leq e^{C_0|\omega|}.$$

This shows that the infinite product converges uniformly on any compact set, and converges absolutely. It is far from showing that $\hat{\phi}(\omega) \in L^2$. We clearly can find much better estimates. For example if $H(\omega)$ satisfies the double-shift orthogonality relations $|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1$ then $|H(\omega)| \leq 1$, which implies that $|\hat{\phi}(\omega)| \leq 1$. This is still far from showing that $\hat{\phi}$ decays sufficiently rapidly at ∞ so that it is square integrable, but it suggests that we can find much sharper estimates.

The following ingenious argument by Daubechies improves the exponential upper bound on $\hat{\phi}(\omega)$ to a polynomial bound. First, let $C_1 \geq 1$ be an upper bound for $|H(\omega)|$,

$$|H(\omega)| \leq C_1 \quad \text{for } 0 \leq \omega < \pi.$$

(Here we do not assume that $H(\omega)$ satisfies double-shift orthogonality.) Set

$$\Phi_k(\omega) = \Pi_{j=1}^k H(\frac{\omega}{2^j}),$$

and note that $|\Phi_k(\omega)| \leq e^{C_0}$ for $|\omega| \leq 1$. Now we bound $|\Phi_k(\omega)|$ for $|\omega| > 1$. For each $|\omega| > 1$ we can uniquely determine the positive integer $K(\omega)$ so that $2^{K-1} < |\omega| \leq 2^K$. Now we derive upper bounds for $|\Phi_k(\omega)|$ in two cases: $k \leq K$ and $k > K$. For $k \leq K$ we have

$$|\Phi_k(\omega)| = \Pi_{j=1}^k |H(\frac{\omega}{2^j})| \leq C_1^K \leq C_1^{1+\log_2|\omega|} = C_1|\omega|^{\log_2 C_1},$$

since $\log_2 C_1^{\log_2|\omega|} = \log_2|\omega|^{\log_2 C_1} = \log_2 C_1 \log_2|\omega|$. For $k > K$ we have

$$|\Phi_k(\omega)| = \Pi_{j=1}^K |H(\frac{\omega}{2^j})| \Pi_{j=1}^{k-K} |H(\frac{\omega}{2^{K+j}})| \leq C_1^K |\Phi_{k-K}(\frac{\omega}{2^K})| \leq C_1|\omega|^{\log_2 C_1} e^{C_0}.$$

Combining these estimates we obtain the uniform upper bound

$$|\Phi_k(\omega)| \leq C_1 e^{C_0}(1 + |\omega|^{\log_2 C_1})$$

for all ω and all integers k. Now when we go to the limit we have the polynomial upper bound

$$|\hat{\phi}(\omega)| \leq C_1 e^{C_0}(1 + |\omega|^{\log_2 C_1}). \tag{8.25}$$

Thus still doesn't give L^2 convergence or smoothness results, but together with information about p zeros of $H(\omega)$ at $\omega = \pi$ we can use it to obtain such results.

The following result clarifies the relationship between the smoothness of the scaling function $\phi(t)$ and the rate of decay of its Fourier transform $\hat{\phi}(\omega)$.

Lemma 46 *Suppose* $\int_{-\infty}^{\infty} |\hat{\phi}(\omega)|(1+|\omega|)^{n+\alpha} d\omega < \infty$, *where* n *is a non-negative integer and* $0 \leq \alpha < 1$. *Then* $\phi(t)$ *has* n *continuous derivatives, and there exists a positive constant* A *such that* $|\phi^{(n)}(t+h) - \phi^{(n)}(t)| \leq A|h|^{\alpha}$, *uniformly in* t *and* h.

SKETCH OF PROOF: From the assumption, the right-hand side of the formal inverse Fourier relation

$$\phi^{(m)}(t) = \frac{i^m}{2\pi} \int_{-\infty}^{\infty} \hat{\phi}(\omega) \omega^m e^{i\omega t} d\omega$$

converges absolutely for $m = 0, 1, \cdots, n$. It is a straightforward application of the Lebesgue dominated convergence theorem to show that $\phi^{(m)}(t)$ is a continuous function of t for all t. Now for $h > 0$

$$\begin{aligned} \frac{\phi^{(n)}(t+h) - \phi^{(n)}(t)}{h^{\alpha}} &= \frac{i^n}{2\pi} \int_{-\infty}^{\infty} \hat{\phi}(\omega) \omega^n \left(\frac{e^{i\omega(t+h)} - e^{i\omega t}}{h^{\alpha}}\right) d\omega \\ &= \frac{i^{n+1}}{\pi} \int_{-\infty}^{\infty} \hat{\phi}(\omega) \omega^{n+\alpha} e^{i\omega(t+\frac{h}{2})} \frac{\sin \frac{\omega h}{2}}{(\omega h)^{\alpha}} d\omega. \quad (8.26) \end{aligned}$$

Note that the function $\sin \frac{\omega h}{2} / (\omega h)^{\alpha}$ is bounded for all ωh. Thus there exists a constant M such that $|\sin \frac{\omega h}{2} / (\omega h)^{\alpha}| \leq M$ for all ωh. It follows from the hypothesis that the integral on the right-hand side of (8.26) is bounded by a constant A. A slight modification of this argument goes through for $h < 0$. Q.E.D.

NOTE: A function $f(t)$ such that $|f(t+h) - f(t)| \leq A|h|^{\alpha}$ is said to be *Hölder continuous* with modulus of continuity α.

Now let's investigate the influence of p zeros at $\omega = \pi$ of the low pass filter function $H(\omega)$. In analogy with our earlier analysis of the cascade algorithm (8.23) we write $H(\omega) = (\frac{1+e^{-i\omega}}{2})^p H_{-1}(\omega)$. Thus the FIR filter $H_{-1}(\omega)$ still has $H_{-1}(0) = 1$, but it doesn't vanish at $\omega = \pi$. Then the infinite product formula for the scaling function

$$\hat{\phi}(\omega) = \Pi_{j=1}^{\infty} H(\frac{\omega}{2^j}), \quad (8.27)$$

changes to

$$\hat{\phi}(\omega) = \Pi_{j=1}^{\infty} \left(\frac{1}{2} + \frac{1}{2} e^{-i\omega/2^j}\right)^p \Pi_{j=1}^{\infty} H_{-1}(\frac{\omega}{2^j})$$

$$= \left(\frac{1-e^{-i\omega}}{i\omega}\right)^p \hat{\phi}_{-1}(\omega). \tag{8.28}$$

The new factor is the Fourier transform of the box function, raised to the power p. From Lemma 46 we have the upper bound

$$|\hat{\phi}_{-1}(\omega)| \le C_1 e^{C_0}(1+|\omega|^{\log_2 C_1}). \tag{8.29}$$

with

$$C_1 = \begin{cases} \max_{\omega\in[0,2\pi]}|H_{-1}(\omega)| \\ 1. \end{cases}$$

This means that $|\hat{\phi}_{-1}(\omega)|$ decays at least as fast as $|\omega|^{\log_2 C_1}$ for $|\omega| \to \infty$, hence that $|\hat{\phi}(\omega)|$ decays at least as fast as $|\omega|^{\log_2 C_1 - p}$ for $|\omega| \to \infty$.

EXAMPLE: Daubechies D_4. The low pass filter is

$$H(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^2 \left[\frac{1}{2}[1+\sqrt{3}] + \frac{1}{2}[1-\sqrt{3}]e^{-i\omega}\right] = \left(\frac{1+e^{-i\omega}}{2}\right)^2 H_{-1}(\omega).$$

Here $p = 2$ and the maximum value of $|H_{-1}(\omega)|$ is $C_1 = \sqrt{3}$ at $\omega = \pi$. Thus $\log_2\sqrt{3} = 0.79248\ldots$ and $|\hat{\phi}(\omega)|$ decays at least as fast as $|\omega|^{-1.207\ldots}$ for $|\omega| \to \infty$. Thus we can apply Lemma 46 with $n = 0$ to show that the Daubechies $D4$ scaling function is continuous with modulus of continuity at least $\alpha = 0.207\ldots$.

We can also use the previous estimate and the computation of C_1 to get a (crude) upper bound for the eigenvalues of the T matrix associated with $H_{-1}(\omega)$, hence for the non-obvious eigenvalues of T_{2N-1} associated with $H(\omega)$. If λ is an eigenvalue of the T matrix associated with $H_{-1}(\omega)$ and $X(\omega)$ is the corresponding eigenvector, then the eigenvalue equation

$$\lambda X(\omega) = |H_{-1}(\frac{\omega}{2})|^2 X(\frac{\omega}{2}) + |H_{-1}(\frac{\omega}{2}+\pi)|^2 X(\frac{\omega}{2}+\pi),$$

is satisfied, where $X(\omega) = \sum_{n=-N+2p}^{N-2p} \mathbf{x}(n)e^{-in\omega}$ and $\mathbf{x}(n)$ is a $2(N-2p)-1$-tuple. We normalize X by requiring that $||X|| = 1$. Let $M = \max_{\omega\in[0,2\pi]}|X(\omega)| = |X(\omega_0)|$. Setting $\omega = \omega_0$ in the eigenvalue equation, and taking absolute values we obtain

$$|\lambda|\cdot M = \left||H_{-1}(\frac{\omega_0}{2})|^2 X(\frac{\omega_0}{2}) + |H_{-1}(\frac{\omega_0}{2}+\pi)|^2 X(\frac{\omega_0}{2}+\pi)\right|$$

$$\le C_1^2 M + C_1^2 M = 2C_1^2 M.$$

Thus $|\lambda| \leq 2C_1^2$ for all eigenvalues associated with $H_{-1}(\omega)$. This means that the non-obvious eigenvalues of T_{2N-1} associated with $H(\omega)$ must satisfy the inequality $|\lambda| \leq 2C_1^2/4^p$. In the case of D_4 where $C_1^2 = 3$ and $p = 2$ we get the upper bound $3/8$ for the non-obvious eigenvalue. In fact, the eigenvalue is $1/4$.

Chapter 9

Other Topics

9.1 The Windowed Fourier transform and the Wavelet Transform

In this and the next section we introduce and study two new procedures for the analysis of time-dependent signals, locally in both frequency and time. The first procedure, the "windowed Fourier transform" is associated with classical Fourier analysis while the second, the (continuous) "wavelet transform" is associated with scaling concepts related to discrete wavelets.

Let $g \in L^2[-\infty, \infty]$ with $||g|| = 1$ and define the time-frequency translation of g by

$$g^{[x_1,x_2]}(t) = e^{2\pi i t x_2} g(t + x_1). \tag{9.1}$$

Now suppose g is centered about the point (t_0, ω_0) in phase (time-frequency) space, i.e., suppose

$$\int_{-\infty}^{\infty} t|g(t)|^2 dt = t_0, \quad \int_{-\infty}^{\infty} \omega|\hat{g}(\omega)|^2 d\omega = \omega_0$$

where $\tilde{g}(\omega) = \int_{-\infty}^{\infty} g(t)e^{-2\pi i\omega t}dt$ is the Fourier transform of $g(t)$. (Note the change in normalization of the Fourier transform for this chapter only.) Then

$$\int_{-\infty}^{\infty} t|g^{[x_1,x_2]}(t)|^2 dt = t_0 - x_1, \quad \int_{-\infty}^{\infty} \omega|\tilde{g}^{[x_1,x_2]}(\omega)|^2 d\omega = \omega_0 + x_2$$

so $g^{[x_1,x_2]}$ is centered about $(t_0 - x_1, \omega_0 + x_2)$ in phase space. To analyze an arbitrary function $f(t)$ in $L^2[-\infty, \infty]$ we compute the inner product

$$F(x_1, x_2) = \langle f, g^{[x_1,x_2]} \rangle = \int_{-\infty}^{\infty} f(t)\overline{g^{[x_1,x_2]}(t)}dt$$

with the idea that $F(x_1, x_2)$ is sampling the behavior of f in a neighborhood of the point $(t_0 - x_1, \omega_0 + x_2)$ in phase space.

As x_1, x_2 range over all real numbers the samples $F(x_1, x_2)$ give us enough information to reconstruct $f(t)$. It is easy to show this directly for functions f such that $f(t)\overline{g}(t-s) \in L^2[-\infty, \infty]$ for all s. Indeed let's relate the windowed Fourier transform to the usual Fourier transform of f (rescaled for this chapter):

$$\tilde{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-2\pi i\omega t}dt, \qquad f(t) = \int_{-\infty}^{\infty} \tilde{f}(\omega)e^{2\pi i\omega t}d\omega. \tag{9.2}$$

Thus since

$$F(x_1, x_2) = \int_{-\infty}^{\infty} f(t)\overline{g(t+x_1)}e^{-2\pi i t x_2}dt,$$

we have

$$f(t)\overline{g(t+x_1)} = \int_{-\infty}^{\infty} F(x_1, x_2)e^{2\pi i t x_2}dx_2.$$

Multiplying both sides of this equation by $g(t+x_1)$ and integrating over x_1 we obtain

$$f(t) = \frac{1}{||g||^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x_1, x_2)g(t+x_1)e^{2\pi i t x_2}dx_1dx_2. \tag{9.3}$$

This shows us how to recover $f(t)$ from the windowed Fourier transform, if f and g decay sufficiently rapidly at ∞. A deeper but more difficult result is the following theorem.

Theorem 62 *The functions $g^{[x_1,x_2]}$ are dense in $L^2[-\infty, \infty]$ as $[x_1, x_2]$ runs over R^2.*

PROOF (Technical): Let V be the closed subspace of $L^2[-\infty, \infty]$ generated by all linear combinations of the functions $g^{[x_1,x_2]}$. Then $L^2[-\infty, \infty] = V \oplus V^\perp$. Then any $h \in L^2[-\infty, \infty]$ can be written uniquely as $h = h_1 + h_2$ with $h_1 \in V$ and $h_2 \in V^\perp$. Let $\mathbf{L} : L^2[-\infty, \infty] \to L^2[-\infty, \infty]$ be the *projection operator* of $L^2[-\infty, \infty]$ on V , i.e., $\mathbf{L}h = h_1$. Our aim is to show that $\mathbf{L} = \mathbf{E}$, the identity operator, so that $L^2[-\infty, \infty] = V$. We will present the basic ideas of the proof, omitting some of the technical details.

Since V is the closure of the space spanned by all finite linear combinations $\sum_i \alpha_i g^{[x_1^{(i)},x_2^{(i)}]}(t)$ it follows that if $h_1(t) \in V$ then $e^{ibt}h_1(t) \in V$ for any real b. Further $(e^{ibt}h_2(t), j(t)) = (h_2(t), e^{-ibt}j(t)) = 0$ for any $j \in V$, so $e^{ibt}h_2(t) \in V^\perp$. Hence $e^{ibt}\mathbf{L}h(t) = e^{ibt}h_1(t) = \mathbf{L}[e^{ibt}]h(t)$, so $\mathbf{L}$ commutes with the operation of multiplication by functions of the form $e^{i\lambda bt}$ for real b. Clearly $\mathbf{L}$ must also commute with multiplication by finite sums of the form $\sum_{b_j} c_j e^{2\pi i \lambda b_j t}$ and, by using the well-known fact that trigonometric polynomials are dense in the space of measurable functions, $\mathbf{L}$ must commute with multiplication by any bounded function $f(t)$ on $(-\infty, \infty)$. Now let Q be a bounded closed interval in $(-\infty, \infty)$ and let χ_Q be the *index function* of Q:

$$\chi_Q(t) = \begin{cases} 1 & \text{if } t \in Q \\ 0 & \text{if } t \notin Q. \end{cases}$$

Consider the function $f_Q = \mathbf{L}\chi_Q$, the projection of χ_Q on V. Since $\chi_Q^2 = \chi_Q$ we have $f_Q(t) = \mathbf{L}\chi_Q(t) = \mathbf{L}\chi_Q^2(t) = \chi_Q(t)\mathbf{L}\chi_Q(t) = \chi_Q(t) f_Q(t)$ so f_Q is nonzero only for $t \in Q$. Furthermore, if Q' is a closed interval with $Q' \subseteq Q$ and $f_{Q'} = \mathbf{L}\chi_{Q'}$ then $f_{Q'}(t) = \mathbf{L}\chi_{Q'}\chi_Q(t) = \chi_{Q'}(t)\mathbf{L}\chi_Q(t) = \chi_{Q'}(t) f_Q(t)$ so $f_{Q'}(t) = f_Q(t)$ for $t \in Q'$ and $f_{Q'}(t) = 0$ for $t \notin Q'$. It follows that there is a unique function $f(t)$ such that $\chi_{\tilde{Q}} f \in L_2(R)$ and $\chi_{\tilde{Q}}(t) f(t) = \mathbf{L}\chi_{\tilde{\mathbf{Q}}}(\mathbf{t})$ for any closed bounded interval $\tilde{Q}$ in $(-\infty, \infty)$. Now let φ be a C^∞ function which is zero in the exterior of $\tilde{Q}$. Then $\mathbf{L}\varphi(t) = \mathbf{L}(\varphi\chi_{\tilde{Q}}(t)) = \varphi(t)\mathbf{L}\chi_{\tilde{Q}}(t) = \varphi(t) f(t)\chi_{\tilde{Q}}(t) = f(t)\varphi(t)$, so $\mathbf{L}$ acts on φ by multiplication by the function $f(t)$. Since as Q runs over all finite subintervals of $(-\infty, \infty)$ the functions φ are dense in $L_2(R)$, it follows that $\mathbf{L} = f(t)\mathbf{E}$.

Now we use the fact that if $h_1(t) \in V$ then so is $h_1(t - a)$ for any real number a. Just as we have shown that $\mathbf{L}$ commutes with multiplication by a function we can show that $\mathbf{L}$ commutes with translations, i.e., $\mathbf{S}_a\mathbf{L} = \mathbf{L}\mathbf{S}_a$ for all translation operators $\mathbf{S}_a h(t) = h(t - a)$. Thus $f(t)h(t - a) = f(t - a)h(t - a)$ for all a and for all $h \in L^2$. Thus $f(t) = f(t - a)$ almost everywhere, which implies that $f(t)$ is a constant. Since $\mathbf{L}^2 = \mathbf{L}$, this constant must be 1. Q.E.D.

Now we see that a general $f \in L^2$ is uniquely determined by the inner products $\langle f, g^{[x_1,x_2]}\rangle$, $-\infty < x_1, x_2 < \infty$. (Suppose $\langle f_1, g^{[x_1,x_2]}\rangle = \langle f_2, g^{[x_1,x_2]}\rangle$ for $f_1, f_2 \in L^2[-\infty, \infty]$ and all x_1, x_2. Then with $f = f_1 - f_2$ we have $\langle f, g^{[x_1,x_2]}\rangle \equiv 0$, so f is orthogonal to the closed subspace generated by the $g^{[x_1,x_2]}$. This closed subspace is $L^2[-\infty, \infty]$ itself. Hence $f = 0$ and $f_1 = f_2$.)

However, as we shall see,the set of basis states $g^{[x_1,x_2]}$ is overcomplete: the coefficients $\langle f, g^{[x_1,x_2]}\rangle$ are not independent of one another, i.e., in general there is no $f \in L^2[-\infty, \infty]$ such that $\langle f, g^{[x_1,x_2]}\rangle = F(x_1, x_2)$ for an arbitrary $F \in L^2(R^2)$. The $g^{[x_1,x_2]}$ are examples of *coherent states*, continuous overcomplete Hilbert space bases which are of interest in quantum optics, quantum field theory, group representation theory, etc. (The situation is analogous to the case of band limited signals. As the Shannon sampling theorem shows, the representation of a band limited signal $f(t)$ in the time domain is overcomplete. The discrete samples $f(\frac{n\pi}{\Omega})$ for n running over the integers are enough to determine f uniquely.)

As an important example we consider the case $g = \pi^{-1/4}e^{-t^2/2}$. (Here g is essentially its own Fourier transform, so we see that g is centered about $(t_0, \omega_0) = (0, 0)$ in phase space. Thus

$$g^{[x_1,x_2]}(t) = \pi^{-1/4}e^{2\pi itx_2}e^{-(t+x_1)^2/2}$$

is centered about $(-x_1, x_2)$.) This example is known as the *Gabor window*. There are two features of the foregoing discussion that are worth special emphasis. First there is the great flexibility in the coherent function approach due to the fact that the function $g \in L^2[-\infty, \infty]$ can be chosen to fit the problem at hand. Second is the fact that coherent states are always overcomplete. Thus it isn't necessary to compute the inner products $\langle f, g^{[x_1,x_2]}\rangle = F(x_1, x_2)$ for every point in phase space. In the windowed Fourier approach one typically samples F at the lattice points $(x_1, x_2) = (ma, nb)$ where a, b are fixed positive numbers and m, n range over the integers. Here, a, b and $g(t)$ must be chosen so that the map $f \to \{F(ma, nb)\}$ is one-to-one; then f can be recovered from the lattice point values $F(ma, nb)$.

Example 10 *Given the function*

$$g(t) = \begin{cases} 1, & |t| \leq \frac{1}{2} \\ 0, & |t| \geq \frac{1}{2}, \end{cases}$$

the set $\{g^{[m,n]}\}$ is an ON basis for $L^2(-\infty, \infty)$. Here, m, n run over the integers. Thus $g^{[x_1,x_2]}$ is overcomplete.

9.1.1 The lattice Hilbert space

There is a new Hilbert space that we shall find particularly useful in the study of windowed Fourier transforms: the *lattice Hilbert space*. This is the space V' of complex valued functions $\varphi(x_1, x_2)$ in the plane R^2 that satisfy the periodicity condition

$$\varphi(a_1 + x_1, a_2 + x_2) = e^{-2\pi ia_1x_2}\varphi(x_1, x_2) \tag{9.4}$$

for $a_1, a_2 = 0, \pm1, \cdots$ and are square integrable over the unit square:

$$\int_0^1\int_0^1 |\varphi(x_1, x_2)|^2 dx_1dx_2 < \infty.$$

The inner product is

$$\langle \varphi_1, \varphi_2 \rangle = \int_0^1 \int_0^1 \varphi(x_1, x_2)\overline{\varphi}(x_1, x_2)dx_1 dx_2.$$

Note that each function $\varphi(x_1, x_2)$ is uniquely determined by its values in the square $\{(x_1, x_2) : 0 \leq x_1, x_2 < 1\}$. It is periodic in x_2 with period 1 and satisfies the "twist" property $\varphi(x_1 + 1, x_2) = e^{-2\pi i x_2}\varphi(x_1, x_2)$.

We can relate this space to $L^2(R) \equiv L^2[-\infty, \infty]$ via the periodizing operator (Weil-Brezin-Zak isomorphism)

$$\mathbf{P}f(x_1, x_2) = \sum_{n=-\infty}^{\infty} e^{2\pi i n x_2} f(n + x_1) \tag{9.5}$$

which is well defined for any $f \in L_2(R)$ which belongs to the Schwartz space. It is straightforward to verify that $\mathbf{f} = \mathbf{P}f$ satisfies the periodicity condition (9.4), hence $\mathbf{f}$ belongs to V'. Now

$$\begin{aligned} & \langle \mathbf{P}f(\cdot,\cdot), \mathbf{P}f'(\cdot,\cdot)\rangle \\ = & \int_0^1 dx_1 \int_0^1 dx_2 \sum_{m,n=-\infty}^{\infty} e^{2\pi i(n-m)x_2} f(n + x_1)\overline{f'(m + x_1)} \\ = & \int_0^1 dx_1 \sum_{n=-\infty}^{\infty} f(n + x_1)\overline{f'(n + x_1)} = \int_{-\infty}^{\infty} f(t_1)\overline{f'}(t)\, dt \\ = & (f, f') \end{aligned}$$

so $\mathbf{P}$ can be extended to an inner product preserving mapping of $L_2(R)$ into V.

It is clear from the Zak transform that if $\varphi(x_1, x_2) = \mathbf{P}f(x_1, x_2)$ then we can recover $f(x_1)$ by integrating with respect to $x_2 : f(x_1) = \int_0^1 \varphi(x_1, y)dy$. Thus we define the mapping $\mathbf{P}^*$ of V' into $L_2(R)$ by

$$\mathbf{P}^*\varphi(t) = \int_0^1 \varphi(t, y)dy, \quad \varphi \in V'. \tag{9.6}$$

Since $\varphi \in V'$ we have

$$\mathbf{P}^*\varphi(t + a) = \int_0^1 \varphi(t, y)e^{-2\pi i a y}dy = \hat{\varphi}_{-a}(t)$$

for a an integer. (Here $\hat{\varphi}_n(t)$ is the nth Fourier coefficient of $\varphi(t, y)$.) The Parseval formula then yields

$$\int_0^1 |\varphi(t, y)|^2 dy = \sum_{a=-\infty}^{\infty} |\mathbf{P}^*\varphi(t + a)|^2$$

so

$$\langle\varphi,\varphi\rangle = \int_0^1\int_0^1 |\varphi(t,y)|^2 dtdy = \int_0^1 \sum_{a=-\infty}^{\infty} |\mathbf{P}^*\varphi(t+a)|^2 dt$$

$$= \int_{-\infty}^{\infty} |\mathbf{P}^*\varphi(t)|^2 dt = (\mathbf{P}^*\varphi, \mathbf{P}^*\varphi).$$

and $\mathbf{P}^*$ is an inner product preserving mapping of V' into $L_2(R)$. Moreover, it is easy to verify that

$$\langle\mathbf{P}f,\varphi\rangle = (f,\mathbf{P}^*\varphi)$$

for $f \in L_2(R)$, $\varphi \in V'$, i.e., $\mathbf{P}^*$ is the adjoint of $\mathbf{P}$. Since $\mathbf{P}^*\mathbf{P} = \mathbf{E}$ on $L_2(R)$ it follows that $\mathbf{P}$ is a unitary operator mapping $L_2(R)$ **onto** V' and $\mathbf{P}^* = \mathbf{P}^{-1}$ is a unitary operator mapping V' **onto** $L_2(R)$.

9.1.2 More on the Zak transform

The Weil-Brezin transform (earlier used in radar theory by Zak, so also called the Zak transform) is very useful in studying the lattice sampling problem for $(f, g^{[x_1,x_2]})$, at the points $(x_1,x_2) = (ma, nb)$ where a, b are fixed positive numbers and m, n range over the integers. This is particularly in the case $a = b = 1$. Restricting to this case for the time being, we let $\psi \in L_2(R)$. Then

$$\psi_{\mathbf{P}}(x_1,x_2) = \mathbf{P}\psi(x_1,x_2,0) = \sum_{k=-\infty}^{\infty} e^{2\pi i k x_2}\psi(x_1+k) \tag{9.7}$$

satisfies

$$\psi_{\mathbf{P}}(k_1+x_1, k_2+x_2) = e^{-2\pi i k_1 x_2}\psi_{\mathbf{P}}(x_1,x_2)$$

for integers k_1, k_2. (Here (9.7) is meaningful if ψ belongs to, say, the Schwartz class. Otherwise $\mathbf{P}\psi = \lim_{n\to s}\mathbf{P}\psi_n$ where $\psi = \lim_{n\to s}\psi_n$ and the ψ_n are Schwartz class functions. The limit is taken with respect to the Hilbert space norm.) If $\psi = g^{[m,n]}(t) = e^{2\pi i t n}g(t+m)$ we have

$$g_{\mathbf{P}}^{[m,n]}(x_1,x_2) = e^{2\pi i(x_1 n - x_2 m)}g_{\mathbf{P}}(x_1,x_2).$$

Thus in the lattice Hilbert space, the functions $g_{\mathbf{P}}^{[m,n]}$ differ from $g_{\mathbf{P}}$ simply by the multiplicative factor $e^{2\pi i(x_1 n - x_2 m)} = \mathbf{E}_{n,m}(x_1,x_2)$, and as n, m range over the integers the $\mathbf{E}_{n,m}$ form an ON basis for the lattice Hilbert space:

$$(\varphi_1,\varphi_2) = \int_0^1\int_0^1 \varphi_1(x_1,x_2)\overline{\varphi_2(x_1,x_2)}dx_1dx_2.$$

Definition 34 *Let $f(t)$ be a function defined on the real line and let $\Xi(t)$ be the characteristic function of the set on which f vanishes:*

$$\Xi(t) = \begin{cases} 1 & \text{if } f(t) = 0 \\ 0 & \text{if } f(t) \neq 0. \end{cases}$$

We say that f is nonzero **almost everywhere** *(a.e.) if the L^2 norm of Ξ is 0, i.e., $||\Xi|| = 0$.*

Thus f is nonzero a.e. provided the Lebesgue integral $\int_{-\infty}^{\infty} \Xi^2(t)dt = 0$, so Ξ belongs to the equivalence class of the zero function. If the support of Ξ is contained in a countable set it will certainly have norm zero; it is also possible that the support is a noncountable set (such as the Cantor set). We will not go into these details here.

Theorem 63 *For $(a, b) = (1, 1)$ and $g \in L^2[-\infty, \infty]$ the transforms $\{g^{[m,n]} : m, n = 0 \pm 1, \pm 2, \cdots\}$ span $L^2[-\infty, \infty]$ if and only if $\mathbf{P}g(x_1, x_2, 0) = g_{\mathbf{P}}(x_1, x_2) \neq 0$ a.e..*

PROOF: Let $\mathcal{M}$ be the closed linear subspace of $L^2[-\infty, \infty]$ spanned by the $\{g^{[m,n]}\}$. Clearly $\mathcal{M} = L^2[-\infty, \infty]$ iff $f = 0$ a.e. is the only solution of $\langle f, g^{[m,n]} \rangle = 0$ for all integers m and n. Applying the Weyl-Brezin -Zak isomorphism $\mathbf{P}$ we have

$$\begin{aligned} \langle f, g^{[m,n]} \rangle &= (\mathbf{P}f, \mathbf{E}_{n,m}\mathbf{P}g) && (9.8) \\ &= ([\mathbf{P}f][\overline{\mathbf{P}g}], \mathbf{E}_{n,m}) = (f_{\mathbf{P}}\bar{g}_{\mathbf{P}}, \mathbf{E}_{n,m}). && (9.9) \end{aligned}$$

Since the functions $\mathbf{E}_{n,m}$ form an ON basis for the lattice Hilbert space it follows that $\langle f, g^{[m,n]} \rangle = 0$ for all integers m, n iff $f_{\mathbf{P}}(x_1, x_2)g_{\mathbf{P}}(x_1, x_2) = 0$, a.e.. If $g_{\mathbf{P}} \neq 0$, a.e. then $f_{\mathbf{P}} = f = 0$ and $\mathcal{M} = L_2(R)$. If $g_{\mathbf{P}} = 0$ on a set S of positive measure on the unit square, and the characteristic function $\chi_S = \mathbf{P}f = f_{\mathbf{P}}$ satisfies $f_{\mathbf{P}}g_{\mathbf{P}} = \chi_S g_{\mathbf{P}} = 0$ a.e., hence $\langle f, g^{[m,n]} \rangle = 0$ and $\mathcal{M} \neq L^2[-\infty, \infty]$. Q.E.D.

In the case $g(t) = \pi^{-1/4}e^{-t^2/2}$ one finds that

$$g_{\mathbf{P}}(x_1, x_2) = \pi^{-1/4} \sum_{k=-\infty}^{\infty} e^{2\pi i k x_2 - (x_1+k)^2/2}.$$

As is well-known, the series defines a Jacobi Theta function. Using complex variable techniques it can be shown (Whittaker and Watson) that this function vanishes at the single point $(\frac{1}{2}, \frac{1}{2})$ in the square $0 \leq x_1 < 1$, $0 \leq x_2 < 1$. Thus $g_{\mathbf{P}} \neq 0$ a.e. and the functions $\{g^{[m,n]}\}$ span $L^2[-\infty, \infty]$. (However, the expansion of an $L^2[-\infty, \infty]$ function in terms of this set is not unique and the $\{g^{[m,n]}\}$ do not form a Riesz basis.)

Corollary 20 *For $(a, b) = (1, 1)$ and $g \in L^2[-\infty, \infty]$ the transforms $\{g^{[m,n]} : m, n = 0, \pm 1, \cdots\}$ form an ON basis for $L^2[-\infty, \infty]$ iff $|g_{\mathbf{P}}(x_1, x_2)| = 1$, a.e.*

PROOF: We have

$$\begin{aligned} \delta_{mm'}\delta_{nn'} &= \langle g^{[m,n]}, g^{[m',n']}\rangle = (E_{n,m}g_{\mathbf{P}}, E_{n',m'}g_{\mathbf{P}}) && (9.10) \\ &= (|g_{\mathbf{P}}|^2, E_{n'-n,m'-m}) && (9.11) \end{aligned}$$

iff $|g_{\mathbf{P}}|^2 = 1$, a.e. Q.E.D.

As an example, let $g = \chi_{[0,1)}$ where

$$\chi_{[0,1)}(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then it is easy to see that $|g_{\mathbf{P}}(x_1, x_2)| \equiv 1$. Thus $\{g^{[m,n]}\}$ is an ON basis for $L_2(R)$.

Theorem 64 *For $(a, b) = (1, 1)$ and $g \in L^2[-\infty, \infty]$, suppose there are constants A, B such that*

$$0 < A \leq |g_{\mathbf{P}}(x_1, x_2)|^2 \leq B < \infty$$

almost everywhere in the square $0 \leq x_1, x_2 < 1$. Then $\{g^{[m,n]}\}$ is a basis for $L_2(R)$, i.e., each $f \in L^2[-\infty, \infty]$ can be expanded **uniquely** *in the form $f = \sum_{m,n} a_{mn}g^{[m,n]}$. Indeed,*

$$a_{mn} = \left(f_{\mathbf{P}}, g_{\mathbf{P}}^{[m,n]}/|g_{\mathbf{P}}|^2\right) = (f_{\mathbf{P}}/g_{\mathbf{P}}, E_{n,m})$$

PROOF: By hypothesis $|g_{\mathbf{P}}|^{-1}$ is a bounded function on the domain $0 \leq x_1, x_2 < 1$. Hence $f_{\mathbf{P}}/g_{\mathbf{P}}$ is square integrable on this domain and, from the periodicity properties of elements in the lattice Hilbert space, $\frac{f_{\mathbf{P}}}{g_{\mathbf{P}}}(x_1 + n, x_2 + m) = \frac{f_{\mathbf{P}}}{g_{\mathbf{P}}}(x_1, x_2)$. It follows that

$$\frac{f_{\mathbf{P}}}{g_{\mathbf{P}}} = \sum a_{mn}E_{n,m}$$

where $a_{mn} = (f_{\mathbf{P}}/g_{\mathbf{P}}, E_{n,m})$, so $f_{\mathbf{P}} = \sum a_{mn}E_{n,m}g_{\mathbf{P}}$. This last expression implies $f = \sum a_{mn}g^{[m,n]}$. Conversely, given $f = \sum a_{mn}g^{[m,n]}$ we can reverse the steps in the preceding argument to obtain $a_{mn} = (f_{\mathbf{P}}/g_{\mathbf{P}}, E_{n,m})$. Q.E.D.

9.1.3 Windowed transforms

The expansion $f = \sum a_{mn} g^{[n,n]}$ is equivalent to the lattice Hilbert space expansion $f_{\mathbf{P}} = \sum a_{mn} E_{n,m} g_{\mathbf{P}}$ or

$$f_{\mathbf{P}} \bar{g}_{\mathbf{P}} = \sum (a_{mn} E_{n,m}) |g_{\mathbf{P}}|^2 \tag{9.12}$$

Now if $g_{\mathbf{P}}$ is a bounded function then $f_{\mathbf{P}} \bar{g}_{\mathbf{P}}(x_1, x_2)$ and $|g_{\mathbf{P}}|^2$ both belong to the lattice Hilbert space and are periodic functions in x_1 and x_2 with period 1. Hence,

$$\begin{aligned} f_{\mathbf{P}} \bar{g}_{\mathbf{P}} &= \sum b_{mn} E_{n,m} && (9.13) \\ |g_{\mathbf{P}}|^2 &= \sum c_{mn} E_{n,m} && (9.14) \end{aligned}$$

with

$$\begin{aligned} b_{mn} &= (f_{\mathbf{P}} \bar{g}_{\mathbf{P}}, E_{n,m}) = (f_{\mathbf{P}}, g_{\mathbf{P}} E_{n,m}) = \langle f, g^{[m,n]} \rangle, && (9.15) \\ c_{mn} &= (g_{\mathbf{P}} \bar{g}_{\mathbf{P}}, E_{n,m}) = \langle g, g^{[m,n]} \rangle. && (9.16) \end{aligned}$$

This gives the Fourier series expansion for $f_{\mathbf{P}} \bar{g}_{\mathbf{P}}$ as the product of two other Fourier series expansions. (We consider the functions f, g, hence $f_{\mathbf{P}}, g_{\mathbf{P}}$ as known.) The Fourier coefficients in the expansions of $f_{\mathbf{P}} \bar{g}_{\mathbf{P}}$ and $|g_{\mathbf{P}}|^2$ are cross-ambiguity functions. If $|g_{\mathbf{P}}|^2$ never vanishes we can solve for the a_{mn} directly:

$$\sum a_{mn} E_{n,m} = (\sum b_{mn} E_{n,m})(\sum c'_{mn} E_{n,m}).$$

where the c'_{mn} are the Fourier coefficients of $|g_{\mathbf{P}}|^{-2}$. However, if $|g_{\mathbf{P}}|^2$ vanishes at some point then the best we can do is obtain the convolution equations $b = a * c$, i.e.,

$$b_{mn} = \sum_{k+k'=m, \ell+\ell'=n} a_{k\ell} c'_{k'\ell'}.$$

(We can approximate the coefficients $a_{k\ell}$ even in the cases where $|g_{\mathbf{P}}|^2$ vanishes at some points. The basic idea is to truncate $\sum a_{mn} E_{n,m}$ to a finite number of nonzero terms and to sample equation (9.12), making sure that $|g_{\mathbf{P}}|(x_1, x_2)$ is nonzero at each sample point. The a_{mn} can then be computed by using the inverse finite Fourier transform.)

The problem of $|g_{\mathbf{P}}|$ vanishing at a point is not confined to an isolated example. Indeed it can be shown that if $g_{\mathbf{P}}$ is an everywhere continuous function in the lattice Hilbert space then it must vanish at at least one point.

9.2 Bases and Frames, Windowed frames

9.2.1 Frames

To understand the nature of the complete sets $\{g^{[m,n]}\}$ it is useful to broaden our perspective and introduce the idea of a **frame** in an arbitrary Hilbert space $\mathcal{H}$. In this more general point of view we are given a sequence $\{\mathbf{f_n}\}$ of elements of $\mathcal{H}$ and we want to find conditions on $\{\mathbf{f_n}\}$ so that we can recover an arbitrary $\mathbf{f} \in \mathcal{H}$ from the inner products $\langle \mathbf{f}, \mathbf{f_n} \rangle$ on $\mathcal{H}$. Let ℓ^2 be the Hilbert space of countable sequences $\{\xi_n\}$ with inner product $(\xi, \eta) = \sum_n \xi_n \bar{\eta}_n$. (A sequence $\{\xi_n\}$ belongs to ℓ^2 provided $\sum_n \xi_n \bar{\xi}_n < \infty$.) Now let $\mathbf{T} : \mathcal{H} \to \ell^2$ be the linear mapping defined by

$$(\mathbf{T}\mathbf{f})_n = \langle \mathbf{f}, \mathbf{f}_n \rangle.$$

We require that $\mathbf{T}$ is a bounded operator from $\mathcal{H}$ to ℓ^2, i.e., that there is a finite $B > 0$ such that $\sum_n |\langle \mathbf{f}, \mathbf{f}_n \rangle|^2 \le B||\mathbf{f}||^2$. In order to recover $\mathbf{f}$ from the $\langle \mathbf{f}, \mathbf{f}_n \rangle$ we want $\mathbf{T}$ to be invertible with $\mathbf{T}^{-1} : \mathcal{R}_\mathbf{T} \to \mathcal{H}$ where $\mathcal{R}_\mathbf{T}$ is the range $\mathbf{T}\mathcal{H}$ of $\mathbf{T}$ in ℓ^2. Moreover, for numerical stability in the computation of $\mathbf{f}$ from the $\langle \mathbf{f}, \mathbf{f}_n \rangle$ we want $\mathbf{T}^{-1}$ to be bounded. (In other words we want to require that a "small" change in the data $\langle \mathbf{f}, \mathbf{f}_n \rangle$ leads to a "small" change in $\mathbf{f}$.) This means that there is a finite $A > 0$ such that $\sum_n |\langle \mathbf{f}, \mathbf{f}_n \rangle|^2 \ge A||\mathbf{f}||^2$. (Note that $\mathbf{T}^{-1}\xi = \mathbf{f}$ if $\xi_n = \langle \mathbf{f}, \mathbf{f}_n \rangle$.) If these conditions are satisfied, i.e., if there exist positive constants A, B such that

$$A||\mathbf{f}||^2 \le \sum_n |\langle \mathbf{f}, \mathbf{f}_n \rangle|^2 \le B||\mathbf{f}||^2$$

for all $\mathbf{f} \in \mathcal{H}$, we say that the sequence $\{\mathbf{f}_n\}$ is a **frame** for $\mathcal{H}$ and that A and B are **frame bounds**. (In general, a frame gives completeness, but also redundancy. There are more terms than the minimal needed to determine $\mathbf{f}$. However, if the set $\{\mathbf{f}_n\}$ is linearly independent, then it forms a basis, called a *Riesz basis*, and there is no redundancy.)

The **adjoint** $\mathbf{T}^*$ of $\mathbf{T}$ is the linear mapping $\mathbf{T}^* : \ell^2 \to \mathcal{H}$ defined by

$$\langle \mathbf{T}^*\xi, \mathbf{f} \rangle = (\xi, \mathbf{T}\mathbf{f})$$

for all $\xi \in \ell^2$, $\mathbf{f} \in \mathcal{H}$. A simple computation yields

$$\mathbf{T}^*\xi = \sum_n \xi_n \mathbf{f}_n.$$

(Since $\mathbf{T}$ is bounded, so is $\mathbf{T}^*$ and the right-hand side is well-defined for all $\xi \in \ell^2$.) Now the bounded self-adjoint operator $\mathbf{S} = \mathbf{T}^*\mathbf{T} : \mathcal{H} \to \mathcal{H}$ is given by

$$\mathbf{S}\mathbf{f} = \mathbf{T}^*\mathbf{T}\mathbf{f} = \sum_n \langle \mathbf{f}, \mathbf{f}_n \rangle \mathbf{f}_n, \tag{9.17}$$

and we can rewrite the defining inequality for the frame as

$$A||\mathbf{f}||^2 \leq \langle \mathbf{T}^*\mathbf{T}\mathbf{f}, \mathbf{f}\rangle \leq B||\mathbf{f}||^2.$$

Since $A > 0$, if $\mathbf{T}^*\mathbf{T}\mathbf{f} = \theta$ then $\mathbf{f} = \theta$, so $\mathbf{S}$ is one-to-one, hence invertible. Furthermore, the range $\mathbf{S}\mathcal{H}$ of $\mathbf{S}$ is $\mathcal{H}$. Indeed, if $\mathbf{S}\mathcal{H}$ is a proper subspace of $\mathcal{H}$ then we can find a nonzero vector $\mathbf{g}$ in $(\mathbf{S}\mathcal{H})^\perp$: $\langle \mathbf{S}\mathbf{f}, \mathbf{g}\rangle = 0$ for all $\mathbf{f} \in \mathcal{H}$. However, $\langle \mathbf{S}\mathbf{f}, \mathbf{g}\rangle = \langle \mathbf{T}^*\mathbf{T}\mathbf{f}, \mathbf{g}\rangle = (\mathbf{T}\mathbf{f}, \mathbf{T}\mathbf{g}) = \sum_n \langle \mathbf{f}, \mathbf{f}_n\rangle\langle \mathbf{f}_n, \mathbf{g}\rangle$. Setting $\mathbf{f} = \mathbf{g}$ we obtain

$$\sum_n |\langle \mathbf{g}, \mathbf{f}_n\rangle|^2 = 0.$$

But then we have $\mathbf{g} = \theta$, a contradiction. thus $\mathbf{S}\mathcal{H} = \mathcal{H}$ and the inverse operator $\mathbf{S}^{-1}$ exists and has domain $\mathcal{H}$.

Since $\mathbf{S}\mathbf{S}^{-1}\mathbf{f} = \mathbf{S}^{-1}\mathbf{S}\mathbf{f} = \mathbf{f}$ for all $\mathbf{f} \in \mathcal{H}$, we immediately obtain two expansions for $\mathbf{f}$ from (9.17):

$$a)\ \mathbf{f} = \sum_{\mathbf{n}} \langle \mathbf{S^{-1}f}, \mathbf{f_n}\rangle \mathbf{f_n} = \sum_{\mathbf{n}} \langle \mathbf{f}, \mathbf{S^{-1}f_n}\rangle \mathbf{f_n} \qquad (9.18)$$

$$b)\ \mathbf{f} = \sum_{\mathbf{n}} \langle \mathbf{f}, \mathbf{f_n}\rangle \mathbf{S^{-1}f_n}. \qquad (9.19)$$

(The second equality in the first expression follows from the identity $\langle \mathbf{S^{-1}f}, \mathbf{f_n}\rangle = \langle \mathbf{f}, \mathbf{S^{-1}f_n}\rangle$, which holds since $\mathbf{S^{-1}}$ is self-adjoint.)

Recall that for a **positive** operator $\mathbf{S}$, i.e., an operator such that $\langle \mathbf{S}\mathbf{f}, \mathbf{f}\rangle \geq \mathbf{0}$ for all $\mathbf{f} \in \mathcal{H}$ the inequalities

$$A||\mathbf{f}||^\mathbf{2} \leq \langle \mathbf{S}\mathbf{f}, \mathbf{f}\rangle \leq \mathbf{B}||\mathbf{f}||^\mathbf{2}$$

for $A, B > 0$ are equivalent to the inequalities

$$A||\mathbf{f}|| \leq ||\mathbf{S}\mathbf{f}|| \leq \mathbf{B}||\mathbf{f}||.$$

This suggests that if the $\{\mathbf{f_n}\}$ form a frame then so do the $\{\mathbf{S^{-1}f_n}\}$.

Theorem 65 *Suppose* $\{\mathbf{f_n}\}$ *is a frame with frame bounds* A, B *and let* $\mathbf{S} = \mathbf{T}^*\mathbf{T}$. *Then* $\{\mathbf{S^{-1}f_n}\}$ *is also a frame, called the* **dual frame** *of* $\{\mathbf{f_n}\}$, *with frame bounds* B^{-1}, A^{-1}.

PROOF: Setting $\mathbf{f} = \mathbf{S^{-1}g}$ we have $B^{-1}||\mathbf{g}|| \leq ||\mathbf{S^{-1}g}|| \leq \mathbf{A^{-1}}||\mathbf{g}||$. Since $\mathbf{S^{-1}}$ is self-adjoint, this implies $B^{-1}||\mathbf{g}||^\mathbf{2} \leq \langle \mathbf{S^{-1}g}, \mathbf{g}\rangle \leq \mathbf{A^{-1}}||\mathbf{g}||^\mathbf{2}$. Then we have $\mathbf{S^{-1}g} = \sum_\mathbf{n} \langle \mathbf{S^{-1}g}, \mathbf{f_n}\rangle \mathbf{S^{-1}f_n}$ so $\langle \mathbf{S^{-1}g}, \mathbf{g}\rangle = \sum_\mathbf{n} \langle \mathbf{S^{-1}g}, \mathbf{f_n}\rangle\langle \mathbf{S^{-1}f_n}, \mathbf{g}\rangle = \sum_\mathbf{n} |\langle \mathbf{g}, \mathbf{S^{-1}f_n}\rangle|^\mathbf{2}$. Hence $\{\mathbf{S^{-1}f_n}\}$ is a frame with frame bounds B^{-1}, A^{-1}. Q.E.D.

We say that $\{\mathbf{f_n}\}$ is a **tight frame** if $A = B$.

Corollary 21 *If* $\{\mathbf{f_n}\}$ *is a tight frame then every* $\mathbf{f} \in \mathcal{H}$ *can be expanded in the form*

$$\mathbf{f} = \mathbf{A^{-1}} \sum_{\mathbf{n}} \langle \mathbf{f}, \mathbf{f_n} \rangle \mathbf{f_n}.$$

PROOF: Since $\{\mathbf{f_n}\}$ is a tight frame we have $A||\mathbf{f}||^{\mathbf{2}} = \langle \mathbf{Sf}, \mathbf{f} \rangle$ or $\langle (\mathbf{S} - \mathbf{AE})\mathbf{f}, \mathbf{f} \rangle = \mathbf{0}$ where $\mathbf{E}$ is the identity operator $\mathbf{Ef} = \mathbf{f}$. Since $\mathbf{S} - \mathbf{AE}$ is a self-adjoint operator we have $||(\mathbf{S} - \mathbf{AE})\mathbf{f}|| = \mathbf{0}$ for all $\mathbf{f} \in \mathcal{H}$. Thus $\mathbf{S} = \mathbf{AE}$. However, from (7.18), $\mathbf{Sf} = \sum_{\mathbf{n}} \langle \mathbf{f}, \mathbf{f_n} \rangle \mathbf{f_n}$. Q.E.D.

Riesz bases

In this section we will investigate the conditions that a frame must satisfy in order for it to define a Riesz basis, i.e., in order that the set $\{\mathbf{f}_n\}$ be linearly independent. Crucial to this question is the adjoint operator. Recall that the **adjoint** $\mathbf{T}^*$ of $\mathbf{T}$ is the linear mapping $\mathbf{T}^* : \ell^2 \to \mathcal{H}$ defined by

$$\langle \mathbf{T}^* \xi, \mathbf{f} \rangle = (\xi, \mathbf{Tf})$$

for all $\xi \in \ell^2$, $\mathbf{f} \in \mathcal{H}$, so that

$$\mathbf{T}^* \xi = \sum_n \xi_n \mathbf{f}_n.$$

Since $\mathbf{T}$ is bounded, it is a straightforward exercise in functional analysis to show that $\mathbf{T}^*$ is also bounded and that if $||\mathbf{T}||^2 = B$ then $||\mathbf{T}||^2 = ||\mathbf{T}^*||^2 = ||\mathbf{TT}^*|| = ||\mathbf{T}^*\mathbf{T}|| = B$. Furthermore, we know that $\mathbf{T}$ is invertible, as is $\mathbf{T}^*\mathbf{T}$, and if $||\mathbf{T}^{-1}||^2 = A^{-1}$ then $||(\mathbf{T}^*\mathbf{T})^{-1}|| = A^{-1}$. However, this doesn't necessarily imply that $\mathbf{T}^*$ is invertible (though it is invertible when restricted to the range of $\mathbf{T}$). $\mathbf{T}^*$ will fail to be invertible if there is a nonzero $\xi \in \ell^2$ such that $\mathbf{T}^*\xi = \sum_n \xi_n \mathbf{f}_n = 0$. This can happen if and only if the set $\{\mathbf{f}_n\}$ is linearly dependent. If $\mathbf{T}^*$ is invertible, then it follows easily that the inverse is bounded and $||(\mathbf{T}^*)^{-1}||^2 = A^{-1}$.

The key to all of this is the operator $\mathbf{TT}^* : \ell^{\mathbf{2}} \to \mathcal{H}$ with the action

$$(\mathbf{TT}^*\xi)_n = \sum_m \xi_m \langle \mathbf{f}_m, \mathbf{f}_n \rangle, \qquad \xi \in \ell^2.$$

Then matrix elements of the infinite matrix corresponding to the operator $\mathbf{TT}^*$ are the inner products $(\mathbf{TT}^*)_{\mathbf{nm}} = \overline{\langle \mathbf{f_n}, \mathbf{f_m} \rangle}$. This is a self-adjoint matrix. If its eigenvalues are all positive and bounded away from zero, with lower bound $A > 0$ then it follows that $\mathbf{T}^*$ is invertible and $||\mathbf{T}^{*-1}||^2 = A^{-1}$. In this case the $\mathbf{f}_n$ form a Riesz basis with Riesz constants A, B.

We will return to this issue when we study biorthogonal wavelets.

9.2.2 Frames of $W - H$ type

We can now relate frames with the lattice Hilbert space construction.

Theorem 66 *For $(a, b) = (1, 1)$ and $g \in L_2(R)$, we have*

$$0 < A \leq |g_{\mathbf{P}}(x_1, x_2)|^2 \leq B < \infty \tag{9.20}$$

almost everywhere in the square $0 \leq x_1, x_2 < 1$ iff $\{g^{[m,n]}\}$ is a frame for $L^2[-\infty, \infty]$ with frame bounds A, B. (By Theorem 64 this frame is actually a basis for $L^2[-\infty, \infty]$.)

PROOF: If (9.20) holds then $g_{\mathbf{P}}$ is a bounded function on the square. Hence for any $f \in L_2(R)$, $f_{\mathbf{P}}\bar{g}_{\mathbf{P}}$ is a periodic function, in x_1, x_2 on the square. Thus

$$\sum_{m,n=-\infty}^{\infty} |\langle f, g^{[m,n]}\rangle|^2 = \sum_{m,n=-\infty}^{\infty} |(f_{\mathbf{P}}, E_{n,m}g_{\mathbf{P}})|^2$$

$$= \sum_{m,n=-\infty}^{\infty} |(f_{\mathbf{P}}\bar{g}_{\mathbf{P}}, E_{n,m})|^2 = ||f_{\mathbf{P}}\bar{g}_{\mathbf{P}}||^2 \tag{9.21}$$

$$= \int_0^1 \int_0^1 |f_{\mathbf{P}}|^2 |g_{\mathbf{P}}|^2 dx_1 dx_2.$$

(Here we have used the Plancherel theorem for the exponentials $E_{n,m}$) It follows from (9.20) that

$$A||f||^2 \leq \sum_{m,n=-\infty}^{\infty} |\langle f, g^{[m,n]}\rangle|^2 \leq B||f||^2, \tag{9.22}$$

so $\{g^{[m,n]}\}$ is a frame.

Conversely, if $\{g^{[m,n]}\}$ is a frame with frame bounds A, B, it follows from (9.22) and the computation (9.21) that

$$A||f_{\mathbf{P}}||^2 \leq \int_0^1 \int_0^1 |f_{\mathbf{P}}|^2 |g_{\mathbf{P}}|^2 dx_1 dx_2 \leq B||f_{\mathbf{P}}||^2$$

for an **arbitrary** $f_{\mathbf{P}}$ in the lattice Hilbert space. (Here we have used the fact that $||f|| = ||f_{\mathbf{P}}||$, since $\mathbf{P}$ is a unitary transformation.) Thus the inequalities (9.20) hold almost everywhere. Q.E.D.

Frames of the form $\{g^{[ma,nb]}\}$ are called **Weyl-Heisenberg** (or **W-H) frames**. The Weyl-Brezin-Zak transform is not so useful for the study of W-H frames with general frame parameters (a, b). (Note from that it is only the product ab that is of significance for the W-H frame parameters. Indeed, the change of variable $t' = t/a$ in (9.1) converts the frame parameters (a, b) to $(a', b') = (1, ab)$.) An easy consequence of the general definition of frames is the following:

Theorem 67 *Let $g \in L_2(R)$ and $a, b, A, B > 0$ such that*

1. $0 < A \leq \sum_m |g(x + ma)|^2 \leq B < \infty$*, a.e.,*

2. g *has support contained in an interval* I *where* I *has length* b^{-1}.

Then the $\{g^{[ma,nb]}\}$ are a W-H frame for $L_2(R)$ with frame bounds $b^{-1}A, b^{-1}B$.

PROOF: For fixed m and arbitrary $f \in L_2(R)$ the function $F_m(t) = f(t)\overline{g(t + ma)}$ has support in the interval $I_m = \{t + ma : x \in I\}$ of length b^{-1}. Thus $F_m(t)$ can be expanded in a Fourier series with respect to the basis exponentials $E_{nb}(t) = e^{2\pi ibnt}$ on I_m. Using the Plancherel formula for this expansion we have

$$\begin{aligned}
\sum_{m,n} |\langle f, g^{[ma,nb]}\rangle|^2 &= \sum_{m,n} |\langle F_m, E_{nb}\rangle|^2 \qquad (9.23)\\
&= \frac{1}{b}\sum_m |\langle F_m, F_m\rangle| = \frac{1}{b}\sum_m \int_{I_m} |f(t)|^2 |g(t + ma)|^2 dt\\
&= \frac{1}{b}\int_{-\infty}^{\infty} |f(t)|^2 \sum_m |g(t + ma)|^2 dt.
\end{aligned}$$

¿From property 1) we have then

$$\frac{A}{b}||f||^2 \leq \sum_{m,n} |\langle f, g^{[ma,nb]}\rangle|^2 \leq \frac{B}{b}||f||^2,$$

so $\{g^{[ma,nb]}\}$ is a W-H frame. Q.E.D.

It can be shown that there are no W-H frames with frame parameters (a, b) such that $ab > 1$. For some insight into this case we consider the example $(a, b) = (N, 1), N > 1$, N an integer. Let $g \in L^2[-\infty, \infty]$. There are two distinct possibilities:

1. There is a constant $A > 0$ such that $A \leq |g_{\mathbf{P}}(x_1, x_2)|$ almost everywhere.

2. There is no such $A > 0$.

Let $\mathcal{M}$ be the closed subspace of $L^2[-\infty,\infty]$ spanned by the functions $\{g^{[mN,n]}, m, n = 0 \pm 1, \pm 2, \cdots\}$ and suppose $f \in L^2[-\infty,\infty]$. Then

$$\langle f, g^{[mN,n]}\rangle = (f_{\mathbf{P}}, E_{n,mN}g_{\mathbf{P}}) = (f_{\mathbf{P}}\bar{g}_{\mathbf{P}}, E_{n,mN}).$$

If possibility 1) holds, we set $f_{\mathbf{P}} = \bar{g}_{\mathbf{P}}^{-1}E_{n_0,1}$. Then $f_{\mathbf{P}}$ belongs to the lattice Hilbert space and $0 = (E_{n_0,1}, E_{n,mN}) = (f_{\mathbf{P}}\bar{g}_{\mathbf{P}}E_{n,mN}) = \langle f, g^{[mN,n]}\rangle$ so $f \in \mathcal{M}^{\perp}$ and $\{g^{[mN,n]}\}$ is not a frame. Now suppose possibility 2) holds. Then according to the proof of Theorem 66, g cannot generate a frame $\{g^{[m,n]}\}$ with frame parameters $(1,1)$ because there is no $A > 0$ such that $A||f||^2 < \sum_{m,n} |\langle f, g^{[m,n]}\rangle|^2$. Since the $\{g^{[mN,n]}\}$ corresponding to frame parameters $(1,N)$ is a proper subset of $\{g^{[m,n]}\}$, it follows that $\{g^{[mN,n]}\}$ cannot be a frame either.

For frame parameters (a,b) with $0 < ab < 1$ it is not difficult to construct W-H frames $\{g^{[ma,nb]}\}$ such that $g \in L_2(R)$ is a smooth function. Taking the case $a = 1, b = \frac{1}{2}$, for example, let v be an infinitely differentiable function on R such that

$$v(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x \geq 1 \end{cases} \tag{9.24}$$

and $0 < v(x) < 1$ if $0 < x < 1$. Set

$$g(x) = \begin{cases} 0, & x \leq 0 \\ v(x), & 0 < x < 1 \\ \sqrt{1 - v^2(x-1)}, & 1 \leq x \leq 2 \\ 0, & 2 < x. \end{cases}$$

Then $g \in L^2[-\infty,\infty]$ is infinitely differentiable and with support contained in the interval $[0,2]$. Moreover, $||g||^2 = 1$ and $\sum_n |g(x+m)|^2 \equiv 1$. It follows immediately from Theorem 67 that $\{g^{[m,n/2]}\}$ is a W-H frame with frame bounds $A = B = 2$.

Theorem 68 *Let $f, g \in L_2(R)$ such that $|f_{\mathbf{P}}(x_1,x_2)|$ $|g_{\mathbf{P}}(x_1,x_2)|$ are bounded almost everywhere. Then*

$$\sum_{m,n} |\langle f, g^{[m,n]}\rangle|^2 = \sum_{m,n} \langle f, f^{[m,n]}\rangle \langle g^{[m,n]}, g\rangle.$$

Since $\langle f, g^{[m,n]}\rangle = (f_{\mathbf{P}}, E_{n,m}g_{\mathbf{P}}) = (f_{\mathbf{P}}\bar{g}_{\mathbf{P}}, E_{n,m})$ we have the Fourier series expansion

$$f_P(x_1,x_2)\overline{g_P(x_1,x_2)} = \sum_{m,n} \langle f, g^{[m,n]}\rangle E_{n,m}(x_1,x_2). \tag{9.25}$$

Since $|f_P|, |g_P|$ are bounded, $f_P\bar{g}_P$ is square integrable with respect to the measure dx_1dx_2 on the square $0 \leq x_1, x_2 < 1$. From the Plancherel formula for double Fourier series, we obtain the identity

$$\int_0^1 \int_0^1 |f_{\mathbf{P}}|^2|g_{\mathbf{P}}|^2 dx_1dx_2 = \sum_{m,n} |\langle f, g^{[m,n]}\rangle|^2.$$

Similarly, we can obtain expansions of the form (9.25) for $f_{\mathbf{P}}\bar{f}_{\mathbf{P}}$ and $g_{\mathbf{P}}\bar{g}_{\mathbf{P}}$. Applying the Plancherel formula to these two functions we find

$$\int_0^1 \int_0^1 |f_{\mathbf{P}}|^2|g_{\mathbf{P}}|^2 dx_1dx_2 = \sum_{m,n} \langle f, f^{[m,n]}\rangle\langle g^{[m,n]}, g\rangle.$$

Q.E.D.

9.2.3 Continuous Wavelets

Here we work out the analog for wavelets of the windowed Fourier transform. Let $w \in L^2[-\infty, \infty]$ with $||w|| = 1$ and define the affine translation of w by

$$w^{(a,b)}(t) = |a|^{-1/2} w\left(\frac{t-b}{a}\right)$$

where $a > 0$. Let $f(t) \in L^2[-\infty, \infty]$. The *integral wavelet transform* of f is the function

$$F_w(a,b) = |a|^{-1/2} \int_{-\infty}^{\infty} f(t)\overline{w}\left(\frac{t-b}{a}\right)dt = (f, w^{(a,b)}). \qquad (9.26)$$

(Note that we can also write the transform as

$$F_w(a,b) = |a|^{1/2} \int_{-\infty}^{\infty} f(au+b)\overline{w}(u)du,$$

which is well-defined as $a \to 0$.) This transform is defined in the time-scale plane. The parameter b is associated with time samples, but frequency sampling has been replaced by the scaling parameter a. In analogy with the windowed Fourier transform, one might expect that the functions $w^{(a,b)}(t)$ span $L_2(R)$ as a, b range over all possible values. However, in general this is not the case. Indeed $L_2(R) = \mathcal{H}^+ \oplus \mathcal{H}^-$ where $\mathcal{H}^+$ consists of the functions f_+ such that the Fourier transform $\mathcal{F}f_+(\omega)$ has support on the positive ω-axis and the functions f_- in $\mathcal{H}^-$

have Fourier transform with support on the negative ω-axis. If the support of $\hat{w}(\omega)$ is contained on the positive ω-axis then the same will be true of $\hat{w}^{(a,b)}(\omega)$ for all $a > 0, b$ as one can easily check. Thus the functions $\{w^{(a,b)}\}$ will not necessarily span $L_2(R)$, though for some choices of w we will still get a spanning set. However, if we choose two nonzero functions $w_\pm \in \mathcal{H}^\pm$ then the (orthogonal) functions $\{w_+^{(a,b)}, w_-^{(a,b)} : a > 0\}$ *will* span $L_2(R)$.

An alternative way to proceed, and the way that we shall follow first, is to compute the samples for all (a, b) such that $a \neq 0$, i.e., also to compute (9.26) for $a < 0$. Now, for example, if the Fourier transform of w has support on the positive ω-axis, we see that, for $a < 0$, $\hat{w}^{(a,b)}(\omega)$ has support on the negative ω-axis. Then it it isn't difficult to show that, indeed, the functions $w^{(a,b)}(t)$ span $L_2(R)$ as a, b range over all possible values (including negative a.). However, to get a convenient inversion formula, we will further require the condition (9.27) to follow (which is just $\hat{w}(0) = 0$).

We will soon see that in order to invert (9.26) and synthesize f from the transform of a single mother wavelet w we shall need to require that

$$\int_{-\infty}^{\infty} w(t)dt = 0. \tag{9.27}$$

Further, we require that $w(t)$ has exponential decay at ∞, i.e., $|w(t)| \leq Ke^{-k|t|}$ for some $k, K > 0$ and all t. Among other things this implies that $|\hat{w}(\omega)|$ is uniformly bounded in ω. Then there is a Plancherel formula.

Theorem 69 *Let $f, g \in L^2[-\infty, \infty]$ and $C = 2\pi \int |\hat{w}(\omega)|^2 \frac{d\omega}{|\omega|}$. Then*

$$C \int_{-\infty}^{\infty} f(t)\overline{g}(t)dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_w(a,b)\overline{G}_w(a,b)\frac{da\, db}{a^2}. \tag{9.28}$$

PROOF: Assume first that $\hat{f}$ and $\hat{g}$ have their support contained in $|\omega| \leq \Omega$ for some finite Ω. Note that the right-hand side of (9.26), considered as a function of b, is the convolution of $f(t)$ and $|a|^{-1/2}\overline{w}(-t/a)$. Thus the Fourier transform of $F_w(a,\cdot)$ is $|a|^{1/2}\hat{f}(\omega)\hat{w}(a\omega)$. Similarly the Fourier transform of $\overline{G}_w(a,\cdot)$ is $|a|^{1/2}\overline{\hat{g}}(\omega)\overline{\hat{w}}(a\omega)$. The standard Plancherel identity gives

$$\int_{-\infty}^{\infty} F_w(a,b)\overline{G}_w(a,b)db = 2\pi \int_{-\infty}^{\infty} |a|\hat{f}(\omega)\overline{\hat{g}}(\omega)|\hat{w}(a\omega)|^2 d\omega.$$

Note that the integral on the right-hand side converges, because f, g are band-limited functions. Multiplying both sides by $da/|a|^2$ and integrating with respect to a and switching the order of integration on the right (justified because the functions are absolutely integrable) we obtain

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_w(a,b)\overline{G}_w(a,b)\frac{da\ db}{a^2} = 2\pi C\int_{-\infty}^{\infty}\hat{f}(\omega)\overline{\hat{g}}(\omega)d\omega.$$

Using the Plancherel formula for Fourier transforms, we have the stated result for band-limited functions.

The rest of the proof is “standard abstract nonsense”. We need to limit the band-limited restriction on f and g. Let f, g be arbitrary L^2 functions and let

$$\hat{f}_N(\omega) = \begin{cases} \hat{f}(\omega) & \text{if } |\omega| \leq N \\ 0 & \text{otherwise,} \end{cases}$$

where N is a positive integer. Then $\hat{f}_N \to \hat{f}$ (in the frequency domain L^2 norm) as $N \to +\infty$, with a similar statement for $\hat{g}_N$. From the Plancherel identity we then have $f_n \to f, g_N \to g$ (in the time domain L^2 norm). Since

$$C\int_{-\infty}^{\infty}|f_N(t) - f_M(t)|^2 dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|F_{w,N}(a,b) - F_{w,M}(a,b)|^2\frac{da\ db}{a^2}$$

it follows easily that $\{F_{w,N}\}$ is a Cauchy sequence in the Hilbert space of square integrable functions in R^2 with measure $da\ db/|a|^2$, and $F_{w,N} \to F_w$ in the norm, as $N \to \infty$. Since the inner products are continuous with respect to this limit, we get the general result. Q.E.D.

You should verify that the requirement $\int_{-\infty}^{\infty} w(t)dt = 0$ ensures that C is finite. At first glance, it would appear that the integral for C diverges.

The synthesis equation for continuous wavelets is as follows.

Theorem 70

$$f(t) = \frac{1}{C}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_w(a,b)|a|^{-1/2}\overline{w}(\frac{t-b}{a})\frac{db\ da}{a^2}. \qquad (9.29)$$

PROOF: Consider the b-integral on the right-hand side of equation (9.29). By the Plancherel formula this can be recast as $2\pi\int_{-\infty}^{\infty}\hat{F}_w(\omega)\overline{\hat{w}}^{(a,\cdot)}(\omega)d\omega$ where the Fourier transform of $F_w(a,\cdot)$ is $|a|^{1/2}\hat{f}(\omega)\hat{w}(a\omega)$, and the Fourier transform of $|a|^{-1/2}w(\frac{t-\cdot}{a})$ is $|a|^{1/2}e^{-it\omega}\hat{w}(a\omega)$. Thus the expression on the right-hand side of (9.29) becomes

$$\frac{2\pi}{C}\int_{-\infty}^{\infty}\frac{da}{|a|}\int_{-\infty}^{\infty}\hat{f}(\omega)|\hat{w}(a\omega)|^2 e^{it\omega}d\omega$$

$$= \frac{2\pi}{C}\int_{-\infty}^{\infty} \hat{f}(\omega)e^{it\omega}d\omega \int_{-\infty}^{\infty} |\hat{w}(a\omega)|^2 \frac{da}{|a|}.$$

The a-integral is just $C/2\pi$, so from the inverse Fourier transform, we see that the expression equals $f(t)$ (provided it meets conditions for pointwise convergence of the inverse Fourier transform). Q.E.D.

Now let's see have to modify these results for the case where we require $a > 0$. We choose two nonzero functions $w_\pm \in \mathcal{H}^\pm$, i.e., w_+ is a positive frequency probe function and w_- is a negative frequency probe. To get a convenient inversion formula, we further require the condition (9.30) (which is just $\hat{w}_+(0) = 0, \hat{w}_-(0) = 0$):

$$\int_{-\infty}^{\infty} w_+(t)dt = \int_{-\infty}^{\infty} w_-(t)dt = 0. \tag{9.30}$$

Further, we require that $w_\pm(t)$ have exponential decay at ∞, i.e., $|w_\pm(t)| \leq Ke^{-k|t|}$ for some $k, K > 0$ and all t. Among other things this implies that $|\hat{w}_\pm(\omega)|$ are uniformly bounded in ω. Finally, we adjust the relative normalization of W_+ and w_- so that

$$C = 2\pi \int_0^{\infty} |\hat{w}_+(\omega)|^2 \frac{d\omega}{|\omega|} = 2\pi \int_{-\infty}^{0} |\hat{w}_-(\omega)|^2 \frac{d\omega}{|\omega|}. \tag{9.31}$$

Let $f(t) \in L^2[-\infty, \infty]$. Now the *integral wavelet transform* of f is the pair of functions

$$F_\pm(a,b) = |a|^{-1/2} \int_{-\infty}^{\infty} f(t)\overline{w}_\pm\left(\frac{t-b}{a}\right)dt = (f, w_\pm^{(a,b)}). \tag{9.32}$$

(Note that we can also write the transform pair as

$$F_\pm(a,b) = |a|^{1/2} \int_{-\infty}^{\infty} f(au+b)\overline{w}_\pm(u)du,$$

which is well-defined as $a \to 0$.) Then there is a Plancherel formula.

Theorem 71 *Let $f, g \in L^2[-\infty, \infty]$. Then*

$$C\int_{-\infty}^{\infty} f(t)\overline{g}(t)dt = \int_0^{\infty}\int_{-\infty}^{\infty} \Big(F_+(a,b)\overline{G}_+(a,b) + F_-(a,b)\overline{G}_-(a,b)\Big) \frac{da\,db}{a^2}. \tag{9.33}$$

PROOF: A straightforward modification of our previous proof. Assume first that $\hat{f}$ and $\hat{g}$ have their support contained in $|\omega| \leq \Omega$ for some finite Ω. Note that the right-hand sides of (9.32), considered as functions of b, are the convolutions of $f(t)$ and $|a|^{-1/2}\overline{w}_{\pm}(-t/a)$. Thus the Fourier transform of $F_{\pm}(a,\cdot)$ is $|a|^{1/2}\hat{f}(\omega)\hat{w}_{\pm}(a\omega)$. Similarly the Fourier transforms of $\overline{G}_{\pm}(a,\cdot)$ are $|a|^{1/2}\overline{\hat{g}}(\omega)\overline{\hat{w}_{\pm}(a\omega)}$. The standard Plancherel identity gives

$$\int_{-\infty}^{\infty} F_+(a,b)\overline{G}_+(a,b)db = 2\pi \int_0^{\infty} |a|\hat{f}(\omega)\overline{\hat{g}}(\omega)|\hat{w}_+(a\omega)|^2 d\omega,$$

$$\int_{-\infty}^{\infty} F_-(a,b)\overline{G}_-(a,b)db = 2\pi \int_{-\infty}^{0} |a|\hat{f}(\omega)\overline{\hat{g}}(\omega)|\hat{w}_-(a\omega)|^2 d\omega.$$

Note that the integrals on the right-hand side converge, because f, g are band-limited functions. Multiplying both sides by $da/|a|^2$, integrating with respect to a (from 0 to $+\infty$) and switching the order of integration on the right we obtain

$$\int_0^{\infty}\int_{-\infty}^{\infty} \left(F_+(a,b)\overline{G}_+(a,b) + F_-(a,b)\overline{G}_-(a,b)\right)\frac{da\,db}{a^2} =$$

$$2\pi C\int_0^{\infty} \hat{f}(\omega)\overline{\hat{g}}(\omega)d\omega + 2\pi C\int_{-\infty}^{0} \hat{f}(\omega)\overline{\hat{g}}(\omega)d\omega$$

$$= 2\pi C\int_{-\infty}^{\infty} \hat{f}(\omega)\overline{\hat{g}}(\omega)d\omega.$$

Using the Plancherel formula for Fourier transforms, we have the stated result for band-limited functions.

The rest of the proof is “standard abstract nonsense”, as before. Q.E.D.

Note that the positive frequency data is orthogonal to the negative frequency data.

Corollary 22

$$\int_0^{\infty}\int_{-\infty}^{\infty} F_+(a,b)\overline{G}_-(a,b)\frac{da\,db}{a^2} = 0. \tag{9.34}$$

PROOF: A slight modification of the proof of the theorem. The standard Plancherel identity gives

$$\int_{-\infty}^{\infty} F_+(a,b)\overline{G_-(a,b)}db = 2\pi\int_0^{\infty} |a|\hat{f}(\omega)\overline{\hat{g}(\omega)}\hat{w}_+(a\omega)\overline{\hat{w}_-(a\omega)}d\omega = 0$$

since $\hat{w}_+(\omega)\overline{\hat{w}_-(\omega)} \equiv 0$. Q.E.D.

The modified synthesis equation for continuous wavelets is as follows.

Theorem 72

$$f(t) = \frac{1}{C}\int_0^\infty \int_{-\infty}^\infty |a|^{-1/2}\left(F_+(a,b)\overline{w}_+(\frac{t-b}{a}) + F_-(a,b)\overline{w}_-(\frac{t-b}{a})\right)\frac{db\,da}{a^2}. \tag{9.35}$$

PROOF: Consider the b-integrals on the right-hand side of equations (9.35). By the Plancherel formula they can be recast as $2\pi\int_{-\infty}^\infty \hat F_\pm(\omega)\overline{\hat w_\pm^{(a,\cdot)}}(\omega)d\omega$ where the Fourier transform of $F_\pm(a,\cdot)$ is $|a|^{1/2}\hat f(\omega)\hat w_\pm(a\omega)$, and the Fourier transform of $|a|^{-1/2}w_\pm(\frac{t-\cdot}{a})$ is $|a|^{1/2}e^{-it\omega}\hat w_\pm(a\omega)$. Thus the expressions on the right-hand sides of (9.35) become

$$\frac{2\pi}{C}\int_{-\infty}^\infty \frac{da}{|a|}\int_{-\infty}^\infty \hat f(\omega)|\hat w_\pm(a\omega)|^2 e^{it\omega}d\omega$$

$$= \frac{2\pi}{C}\int_{-\infty}^\infty \hat f(\omega)e^{it\omega}d\omega\int_{-\infty}^\infty |\hat w_\pm(a\omega)|^2\frac{da}{|a|}.$$

Each a-integral is just $C/2\pi$, so from the inverse Fourier transform, we see that the expression equals $f(t)$ (provided it meets conditions for pointwise convergence of the inverse Fourier transform). Q.E.D.

Can we get a continuous transform for the case $a > 0$ that uses a single wavelet? Yes, but not any wavelet will do. A convenient restriction is to require that $w(t)$ is a *real-valued* function with $||w|| = 1$. In that case it is easy to show that $\overline{\hat w(\omega)} = \hat w(-\omega)$, so $|\hat w(\omega)| = |\hat w(-\omega)|$. Now let

$$w_+(t) = \int_0^\infty \hat w(\omega)e^{i\omega t}d\omega, \qquad w_-(t) = \int_{-\infty}^0 \hat w(\omega)e^{i\omega t}d\omega.$$

Note that

$$w(t) = w_+(t) + w_-(t), \qquad ||w_+|| = ||w_-||$$

and that $w_\pm$ are, respectively, positive and negative frequency wavelets. we further require the zero area condition (which is just $\hat w(o) = \hat w_+(0) = 0, \hat w_-(0) = 0$):

$$\int_{-\infty}^\infty w_+(t)dt = \int_{-\infty}^\infty w_-(t)dt = 0, \tag{9.36}$$

and that $w(t)$ have exponential decay at ∞. Then

$$C = 2\pi\int_0^\infty |\hat w(\omega)|^2\frac{d\omega}{|\omega|} = 2\pi\int_{-\infty}^0 |\hat w_\pm(\omega)|^2\frac{d\omega}{|\omega|} \tag{9.37}$$

exists. Let $f(t) \in L^2[-\infty,\infty]$. Here the integral wavelet transform of f is the function

$$F(a,b) = |a|^{-1/2}\int_{-\infty}^{\infty} f(t)w\left(\frac{t-b}{a}\right)dt = (f, w^{(a,b)}). \qquad (9.38)$$

Theorem 73 *Let $f, g \in L^2[-\infty,\infty]$, and $w(t)$ a real-valued wavelet function with the properties listed above. Then*

$$\frac{C}{2}\int_{-\infty}^{\infty} f(t)\overline{g}(t)dt = \int_0^{\infty}\int_{-\infty}^{\infty} F(a,b)\overline{G}(a,b)\frac{da\, db}{a^2}. \qquad (9.39)$$

PROOF: This follows immediately from Theorem 71 and the fact that

$$\int_0^{\infty}\int_{-\infty}^{\infty} F(a,b)\overline{G}(a,b)\frac{da\, db}{a^2} = \int_0^{\infty}\int_{-\infty}^{\infty} (F_+(a,b)+F_-(a,b))\left(\overline{G}_+(a,b)+\overline{G}_-(a,b)\right)\frac{da\, db}{a^2}$$

$$= \int_0^{\infty}\int_{-\infty}^{\infty} \left(F_+(a,b)\overline{G}_+(a,b)+F_-(a,b)\overline{G}_-(a,b)\right)\frac{da\, db}{a^2},$$

due to the orthogonality relation (9.34). Q.E.D.

The synthesis equation for continuous real wavelets is as follows.

Theorem 74

$$f(t) = \frac{2}{C}\int_0^{\infty}\int_{-\infty}^{\infty} |a|^{-1/2}F(a,b)w(\frac{t-b}{a})\frac{db\, da}{a^2}. \qquad (9.40)$$

The continuous wavelet transform is overcomplete, just as is the windowed Fourier transform. To avoid redundancy (and for practical computation where one cannot determine the wavelet transform for a continuum of a, b values) we can restrict attention to discrete lattices in time-scale space. Then the question is which lattices will lead to bases for the Hilbert space. Our work with discrete wavelets in earlier chapters has already given us many nontrivial examples of of a lattice and wavelets that will work. We look for other examples.

9.2.4 Lattices in Time-Scale Space

To define a lattice in the time-scale space we choose two nonzero real numbers $a_0, b_0 > 0$ with $a_0 \neq 1$. Then the lattice points are $a = a_0^m, b = nb_0a_0^m$, $m, n = 0, \pm 1, \cdots$, so

$$w^{mn}(t) = w^{(a_0^m, nb_0a_0^m)}(t) = a_0^{-m/2}w(a_0^{-m}t - nb_0).$$

Note that if w has support contained in an interval of length ℓ then the support of w^{mn} is contained in an interval of length $a_0^{-m}\ell$. Similarly, if $\mathcal{F}w$ has support contained in an interval of length L then the support of $\mathcal{F}w^{mn}$ is contained in an interval of length $a_0^m L$. (Note that this behavior is very different from the behavior of the Heisenberg translates $g^{[ma,nb]}$. In the Heisenberg case the support of g in either position or momentum space is the same as the support of $g^{[ma,nb]}$. In the affine case the sampling of position-momentum space is on a logarithmic scale. There is the possibility, through the choice of m and n, of sampling in smaller and smaller neighborhoods of a fixed point in position space.)

The affine translates $w^{(a,b)}$ are called **continuous wavelets** and the function w is a **mother wavelet**. The map $\mathbf{T}:\mathbf{f}\to\langle\mathbf{f},\mathbf{g_\pm^{mn}}\rangle$ is the **wavelet transform**.

NOTE: This should all look very familiar to you. The lattice $a_0=2^j, b_0=1$ corresponds to the multiresolution analysis that we studied in the preceding chapters.

9.3 Affine Frames

The general definitions and analysis of frames presented earlier clearly apply to wavelets. However, there is no affine analog of the Weil-Brezin-Zak transform which was so useful for Weyl-Heisenberg frames. Nonetheless we can prove the following result directly.

Lemma 47 *Let $w_+\in L_2(R)$ such that the support of $\mathcal{F}g$ is contained in the interval $[\ell,L]$ where $0<\ell<L<\infty$, and let $a_0>1, b_0>0$ with $(L-\ell)b_0\le 1$. Suppose also that*

$$0<A\le\sum_m|\mathcal{F}w_+(a_0^m\omega)|^2\le B<\infty$$

for almost all $\omega\ge 0$. Then $\{w_+^{mn}\}$ is a frame for $\mathcal{H}^+$ with frame bounds $A/b_0, B/b_0$.

PROOF: Let $f\in\mathcal{H}^+$ and note that $w_+\in\mathcal{H}^+$. For fixed m the support of $\mathcal{F}f(a_0^m y)\overline{\mathcal{F}w_+}(\omega)$ is contained in the interval $\ell\le\omega\le\ell+1/b_0$ (of length $1/b_0$). Then

$$\begin{aligned}\sum_{m,n}|\langle f,w_+^{mn}\rangle|^2 &= \sum_{m,n}|\langle\mathcal{F}f,\mathcal{F}g^{mn}\rangle|^2 \qquad (9.41)\\ &= \sum_{m,n}a_0^{-m}|\int_{-0}^{\infty}\mathcal{F}f(a_0^{-m}\omega)\overline{\mathcal{F}w_+}(\omega)e^{-inb_0\omega}d\omega|^2\end{aligned}$$

$$\begin{aligned}
&= \sum_m \frac{a_0^{-m}}{b_0}\int_\ell^{\ell+1/b_0} |\mathcal{F}f(a_0^{-m}\omega)\mathcal{F}w_+(\omega)|^2 d\omega \\
&= \frac{1}{b}\sum_m \int_0^\infty |\mathcal{F}f(\omega)\mathcal{F}w_+(a_0^m\omega)|^2 d\omega \\
&= \frac{1}{b}\int_0^\infty |\mathcal{F}f(\omega)|^2 \left(\sum_m |\mathcal{F}w_+(a_0^m\omega)|^2\right) d\omega.
\end{aligned}$$

Since $||f||^2 = \int_0^\infty |\mathcal{F}f(\omega)|^2 d\omega$ for $f \in \mathcal{H}^+$, the result

$$A||f||^2 \le \sum_{m,n} |\langle f, w_+^{mn}\rangle|^2 \le B||f||^2$$

follows. Q.E.D.

A very similar result characterizes a frame for $\mathcal{H}^-$. (Just let ω run from $-\infty$ to 0.) Furthermore, if $\{w_+^{mn}\}, \{w_-^{mn}\}$ are frames for $\mathcal{H}^+, \mathcal{H}^-$, respectively, corresponding to lattice parameters a_0, b_0, then $\{w_+^{mn}, w_-^{mn}\}$ is a frame for $L_2(R)$

Examples 5 1. *For lattice parameters $a_0 = 2, b_0 = 1$, choose $\hat{w}_+ = \chi_{[1,2)}$ and $\hat{w}_- = \chi_{(-2,-1]}$. Then w_+ generates a tight frame for $\mathcal{H}^+$ with $A = B = 1$ and g_- generates a tight frame for $\mathcal{H}^-$ with $A = B = 1$. Thus $\{w_+^{mn}, w_-^{mn}\}$ is a tight frame for $L_2(R)$. (Indeed, one can verify directly that $\{w_\pm^{mn}\}$ is an ON basis for $L_2(R)$.*

2. *Let w be the function such that*

$$\mathcal{F}w(\omega) = \frac{1}{\sqrt{\ln a}}\begin{cases} 0 & \text{if } \omega \le \ell \\ \sin\frac{\pi}{2}v\left(\frac{\omega-\ell}{\ell(a-1)}\right) & \text{if } \ell < \omega \le a\ell \\ \cos\frac{\pi}{2}v\left(\frac{\omega-a\ell}{a\ell(a-1)}\right) & \text{if } a\ell < \omega \le a^2\ell \\ 0 & \text{if } a^2\ell < \omega \end{cases}$$

where $v(x)$ is defined as in (9.24). Then $\{w^{mn}\}$ is a tight frame for $\mathcal{H}^+$ with $A = B = \frac{1}{b\ln a}$. Furthermore, if $w_+ = w$ and $w_- = \bar{w}$ then $\{w_\pm^{mn}\}$ is a tight frame for $L_2(R)$.

Suppose $w \in L_2(R)$ such that $\mathcal{F}w(\omega)$ is bounded almost everywhere and has support in the interval $\left[-\frac{1}{2b}, \frac{1}{2b}\right]$. Then for any $f \in L_2(R)$ the function

$$a_0^{-m/2}\mathcal{F}f(a_0^{-m}\omega)\overline{\mathcal{F}w(\omega)}$$

has support in this same interval and is square integrable. Thus

$$\begin{aligned}\sum_{m,n}|\langle f, w_{mn}\rangle|^2 &= \sum_{m,n}|a_0^{-m/2}\int_{-\infty}^{\infty}\mathcal{F}f(a_0^{-m}\omega)\overline{\mathcal{F}w(\omega)}e^{-2\pi i\omega b_0}d\omega|^2 \quad (9.42)\\ &= b_0^{-1}\sum_m\int_{-\infty}^{\infty}a_0^{-m}|\mathcal{F}f(a_0^{-m}\omega)\mathcal{F}w(\omega)|^2 d\omega\\ &= \frac{1}{b_0}\int_{-\infty}^{0}|\mathcal{F}f(\omega)|^2\sum_m|\mathcal{F}w(a_0^m\omega)|^2 d\omega\\ &+ \frac{1}{b_0}\int_{0}^{\infty}|\mathcal{F}f(\omega)|^2\sum_m|\mathcal{F}w(a_0^m\omega)|^2 d\omega.\end{aligned}$$

It follows from the computation that if there exist constants $A, B > 0$ such that

$$A \le \sum_m |\mathcal{F}w(a_0^m\omega)|^2 \le B$$

for almost all ω, then the single mother wavelet w generates an affine frame.

Of course, the multiresolution analysis of the preceding chapters provides a wealth of examples of affine frames, particularly those that lead to orthonormal bases. In the next section we will use multiresolution analysis to find affine frames that correspond to biorthogonal bases.

9.4 Biorthogonal Filters and Wavelets

9.4.1 Resumé of Basic Facts on Biorthogonal Filters

Previously, our main emphasis has been on orthogonal filter banks and orthogonal wavelets. Now we will focus on the more general case of biorthogonality. For filter banks this means, essentially, that the analysis filter bank is invertible (but not necessarily unitary) and the synthesis filter bank is the inverse of the analysis filter bank. We recall some of the main facts from Section 6.7, in particular, Theorem 6.7: A 2-channel filter bank gives perfect reconstruction when

$$\begin{aligned}\text{No distortion}: F_0(z)H_0(z) + F_1(z)H_1(z) &= 2z^{-\ell}\\ \text{Alias cancellation}: F_0(z)H_0(-z) + F_1(z)H_1(-z) &= 0. \quad (9.43)\end{aligned}$$

In matrix form this reads

$$[F_0(z) \quad F_1(z)]\left[\begin{array}{cc} H_0(z) & H_0(-z)\\ H_1(z) & H_1(-z)\end{array}\right] = [2z^{-\ell} \quad 0],$$

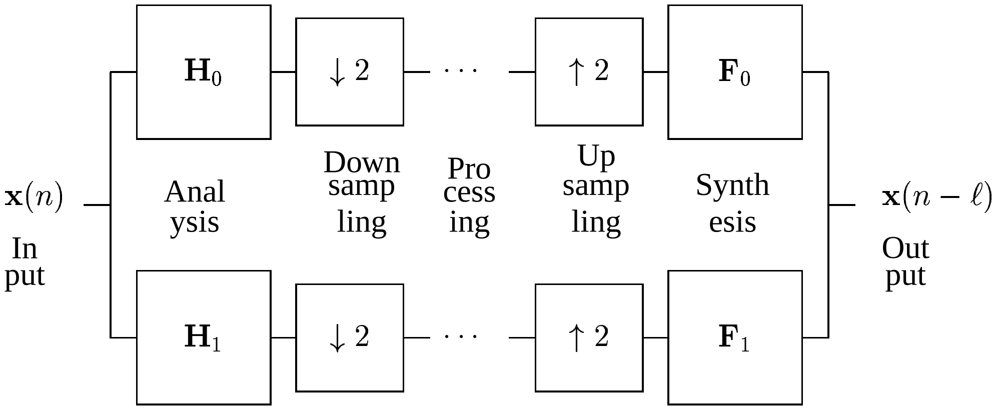

Figure 9.1: Perfect reconstruction 2-channel filter bank

where the 2×2 matrix is the *analysis modulation matrix* $\mathbf{H}_m(z)$.

This is the mathematical expression of Figure 9.1.

We can solve the alias cancellation requirement by defining the synthesis filters in terms of the analysis filters:

$$F_0(z) = H_1(-z), \qquad F_1(z) = -H_0(-z)$$

We introduce the (lowpass) *product filter*

$$P_0(z) = F_0(z)H_0(z)$$

and the (high pass) product filter

$$P_1(z) = F_1(z)H_1(z).$$

¿From our solution of the alias cancellation requirement we have $P_0(z) = H_0(z)H_1(-z)$ and $P_1(z) = -H_0(-z)H_1(z) = -P_0(-z)$. Thus

$$P_0(z) - P_0(-z) = 2z^{-\ell}. \tag{9.44}$$

Note that the even powers of z in $P_0(z)$ cancel out of (9.44). The restriction is only on the odd powers. This also tells us the ℓ is an odd integer. (In particular, it can never be 0.)

The construction of a perfect reconstruction 2-channel filter bank has been reduced to two steps:

1. Design the lowpass filter P_0 satisfying (9.44).
2. Factor P_0 into F_0H_0, and use the alias cancellation solution to get F_1, H_1.

A further simplification involves recentering P_0 to factor out the delay term. Set $P(z) = z^\ell P_0(z)$. Then equation (9.44) becomes the *halfband filter* equation

$$P(z) + P(-z) = 2. \tag{9.45}$$

This equation says the coefficients of the even powers of z in $P(z)$ vanish, except for the constant term, which is 1. The coefficients of the odd powers of z are undetermined design parameters for the filter bank.

In terms of the analysis modulation matrix, and the *synthesis modulation matrix* that will be defined here, the alias cancellation and no distortion conditions read

$$\begin{bmatrix} F_0(z) & F_1(z) \\ F_0(-z) & F_1(-z) \end{bmatrix} \begin{bmatrix} H_0(z) & H_0(-z) \\ H_1(z) & H_1(-z) \end{bmatrix} = \begin{bmatrix} 2z^{-\ell} & 0 \\ o & 2(-z)^{-\ell} \end{bmatrix},$$

where the 2 × 2 F-matrix is the *synthesis modulation matrix* $\mathbf{F}_m(z)$. (Note the transpose distinction between $\mathbf{H}_m(z)$ and $\mathbf{F}_m(z)$.) If we recenter the filters then the matrix condition reads

$$\mathbf{F}_m(z)\dot{\mathbf{H}}_m(z) = 2\mathbf{I}$$

where $\dot{H}_0(z) = z^\ell H_0(z)$, $\dot{H}_0(-z) = (-z)^\ell H_0(-z)$, $\dot{H}_1(z) = z^\ell H_1(z)$, and $\dot{H}_1(-z) = (-z)^\ell H_1(-z)$. (Note that since these are *finite matrices*, the fact that $\dot{\mathbf{H}}_m(z)$ has a left inverse implies that it has the same right inverse, and is invertible.)

Example 11 *Daubechies D_4 half band filter is*

$$P(z) = \frac{1}{16}\left(-z^3 + 9z + 16 + 9z^{-1} - z^{-3}\right).$$

The shifted filter $P_0(z) = z^{-\ell}P(z)$ must be a polynomial in z^{-1} with constant term 1. *Thus $\ell = 3$ and*

$$P_0(z) = \frac{1}{16}\left(-1 + 9z^{-2} + 16z^{-3} + 9z^{-4} - z^{-6}\right).$$

Note that $p = 2$ for this filter, so $P_0(z) = (1 + z^{-1})^4Q(z)$ where Q has only two roots, $c = 2 - \sqrt{3}$ and $\frac{1}{c} = 2 + \sqrt{3}$. There are a variety of factorizations

$P_0(z) = F_0(z)H_0(z)$, depending on which factors are assigned to H_0 and which to F_0. For the construction of filters it makes no difference if the factors are assigned to H_0 or to F_0 (there is no requirement that H_0 be a low pass filter, for example) so we will list the possibilities in terms of Factor 1 and Factor 2.

REMARK: If, however, we want to use these factorizations for the construction of biorthogonal or orthogonal wavelets, then H_0 and F_0 are required to be low-pass. Furthermore it is not enough that the no distortion and alias cancellation conditions hold. The T matrices corresponding to the low pass filters H_0 and F_0 must each have the proper eigenvalue structure to guarantee L^2 convergence of the cascade algorithm. Thus some of the factorizations listed in the table will not yield useful wavelets.

Case	**Factor1**	**DegreeN**	**Factor2**	**DegreeN**
a	1	0	$(1+z^{-1})^4(c-z^{-1})(c^{-1}-z^{-1})$	6
b	$(1+z^{-1})$	1	$(1+z^{-1})^3(c-z^{-1})(c^{-1}-z^{-1})$	5
b'	$(c-z^{-1})$	1	$(1+z^{-1})^4(c^{-1}-z^{-1})$	5
c	$(1+z^{-1})^2$	2	$(1+z^{-1})^2(c-z^{-1})(c^{-1}-z^{-1})$	4
c'	$(1+z^{-1})(c-z^{-1})$	2	$(1+z^{-1})^3(c^{-1}-z^{-1})$	4
c''	$(c-z^{-1})(c^{-1}-z^{-1})$	2	$(1+z^{-1})^4$	4
d	$(1+z^{-1})^3$	3	$(1+z^{-1})(c-z^{-1})(c^{-1}-z^{-1})$	3
d'	$(1+z^{-1})^2(c-z^{-1})$	3	$(1+z^{-1})^2(c^{-1}-z^{-1})$	3

For cases b', c', d' we can switch $c \leftrightarrow c^{-1}$ to get new possibilities. Filters b', c'' are rarely used. A common notation is $(N_A+1)/(N_S+1)$ where N_A is the degree of the analysis filter H_0 and N_S is the degree of the synthesis filter F_0. Thus from c we could produce a $5/3$ filter $H_0(z) = \frac{1}{8}(-1+2z^{-1}+6z^{-2}+2z^{-3}-z^{-4})$ and $F_0(z) = \frac{1}{2}(1+2z^{-1}+z^{-2})$, or a $3/5$ filter with H_0 and F_0 interchanged. The orthonormal Daubechies D_4 filter $(4/4)$ comes from case d'.

Let's investigate what these perfect reconstruction requirements say about the finite impulse response vectors $\mathbf{h}_0(n), \mathbf{h}_1(n), \mathbf{f}_0(n), \mathbf{f}_1(n)$. The half-band filter condition for $P(z)$, recentered , says

$$\sum_n \mathbf{h}_0(\ell+n)\mathbf{f}_0(2k-n) = \delta_{0k}, \tag{9.46}$$

and

$$\sum_n \mathbf{h}_1(\ell+n)\mathbf{f}_1(2k-n) = -\delta_{0k}, \tag{9.47}$$

or

$$\sum_n \breve{\mathbf{h}}_0(n)\mathbf{f}_0(2k+n) = \delta_{0k}, \qquad (9.48)$$

and

$$\sum_n \breve{\mathbf{h}}_1(n)\mathbf{f}_1(2k+n) = -\delta_{0k}, \qquad (9.49)$$

where

$$\breve{\mathbf{h}}_j(n) = \mathbf{h}_j(\ell - n),$$

so $\mathbf{f}_0(n) = (-1)^n\mathbf{h}_1(\ell - n)$, $\mathbf{f}_1(n) = (-1)^{n+1}\mathbf{h}_0(\ell - n)$. The anti-alias conditions

$$\dot{H}_0(z)F_1(z) - \dot{H}_0(-z)F_1(-z) = 0 \qquad \dot{H}_1(z)F_0(z) - \dot{H}_1(-z)F_0(-z) = 0$$

imply

$$\sum_n \breve{\mathbf{h}}_0(n)\mathbf{f}_1(2k+n) = 0, \qquad (9.50)$$

and

$$\sum_n \breve{\mathbf{h}}_1(n)\mathbf{f}_0(2k+n) = 0. \qquad (9.51)$$

Expression (9.48) gives us some insight into the support of $\breve{\mathbf{h}}_0(n)$ and $\mathbf{f}_0(n)$. Since $P_0(z)$ is an even order polynomial in z^{-1} it follows that the sum of the orders of $H_0(z)$ and $F_0(z)$ must be even. This means that $\breve{\mathbf{h}}_0(n)$ and $\mathbf{f}_0(n)$ are each nonzero for an even number of values, or each are nonzero for an odd number of values.

9.4.2 Biorthogonal Wavelets: Multiresolution Structure

In this section we will introduce a multiresolution structure for biorthogonal wavelets, a generalization of what we have done for orthogonal wavelets. Again there will be striking parallels with the study of biorthogonal filter banks. We will go through this material rapidly, because it is so similar to what we have already presented.

Definition 35 *Let $\{V_j : j = \cdots, -1, 0, 1, \cdots\}$ be a sequence of subspaces of $L^2[-\infty, \infty]$ and $\phi \in V_0$. Similarly, let $\{\tilde{V}_j : j = \cdots, -1, 0, 1, \cdots\}$ be a sequence of subspaces of $L^2[-\infty, \infty]$ and $\tilde{\phi} \in \tilde{V}_0$.This is a biorthogonal multiresolution analysis for $L^2[-\infty, \infty]$ provided the following conditions hold:*

1. *The subspaces are nested: $V_j \subset V_{j+1}$ and $\tilde{V}_j \subset \tilde{V}_{j+1}$.*

2. *The union of the subspaces generates L^2 : $\overline{\cup_{j=-\infty}^{\infty} V_j} = \overline{\cup_{j=-\infty}^{\infty} \tilde{V}_j} = L^2[-\infty, \infty]$.*

3. *Separation: $\cap_{j=-\infty}^{\infty} V_j = \cap_{j=-\infty}^{\infty} \tilde{V}_j = \{0\}$, the subspace containing only the zero function. (Thus only the zero function is common to all subspaces V_j, or to all subspaces $\tilde{V}_j$.)*

4. *Scale invariance: $f(t) \in V_j \iff f(2t) \in V_{j+1}$, and $\tilde{f}(t) \in \tilde{V}_j \iff \tilde{f}(2t) \in \tilde{V}_{j+1}$.*

5. *Shift invariance of V_0 and $\tilde{V}_0$: $f(t) \in V_0 \iff f(t-k) \in V_0$ for all integers k, and $\tilde{f}(t) \in \tilde{V}_0 \iff \tilde{f}(t-k) \in \tilde{V}_0$ for all integers k.*

6. *Biorthogonal bases: The set $\{\phi(t-k) : k = 0, \pm 1, \cdots\}$ is a Riesz basis for V_0, the set $\{\tilde{\phi}(t-k) : k = 0, \pm 1, \cdots\}$ is a Riesz basis for $\tilde{V}_0$ and these bases are biorthogonal:*

$$< \phi_{ok}, \tilde{\phi}_{0\ell} >= \int_{-\infty}^{\infty} \phi(t-k)\tilde{\phi}(t-\ell)dt = \delta_{k\ell}.$$

Now we have two scaling functions, the **synthesizing function** *$\phi(t)$, and the* **analyzing function** *$\tilde{\phi}(t)$. The $\tilde{V}, \tilde{W}$ spaces are called the* **analysis multiresolution** *and the spaces V, W are called the* **synthesis multiresolution**.

In analogy with orthogonal multiresolution analysis we can introduce complements W_j of V_j in V_{j+1}, and $\tilde{W}_j$ of $\tilde{V}_j$ in $\tilde{V}_{j+1}$:

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

However, these will no longer be orthogonal complements. We start by constructing a Riesz basis for the analysis wavelet space $\tilde{W}_0$. Since $\tilde{V}_0 \subset \tilde{V}_1$, the analyzing function $\tilde{\phi}(t)$ must satisfy the *analysis dilation equation*

$$\tilde{\phi}(t) = 2\sum_k \breve{\mathbf{h}}_0(k)\tilde{\phi}(2t-k), \tag{9.52}$$

where

$$\sum_k \breve{\mathbf{h}}_0(k) = 1$$

for compatibility with the requirement

$$\int_{-\infty}^{\infty} \tilde{\phi}(t)dt = 1.$$

Similarly, since $V_0 \subset V_1$, the synthesis function $\phi(t)$ must satisfy the *synthesis dilation equation*

$$\phi(t) = 2\sum_k \mathbf{f}_0(k)\phi(2t-k), \tag{9.53}$$

where

$$\sum_k \mathbf{f}_0(k) = 1 \tag{9.54}$$

for compatibility with the requirement

$$\int_{-\infty}^{\infty} \phi(t)dt = 1.$$

REMARK 1: There is a problem here. If the filter coefficients are derived from the half band filter $P_0(z) = H_0(z)F_0(z)$ as in the previous section, then for low pass filters we have $2 = P(1) = H_0(1)F_0(1) = [\sum_k \mathbf{h}_0(k)][\sum_k \mathbf{f_0}(k)]$, so we can't have $\sum_k \mathbf{h}_0(k) = \sum_k \mathbf{f}_0(k) = 1$ simultaneously. To be definite we will always choose the H_0 filter such that $\sum_k \mathbf{h}_0(k) = \sum \breve{\mathbf{h}}_0(k) = 1$. Then we must replace the expected synthesis dilation equations (9.53) and (9.54) by

$$\phi(t) = \sum_k \mathbf{f}_0(k)\phi(2t-k), \tag{9.55}$$

where

$$\sum_k \mathbf{f}_0(k) = 2 \tag{9.56}$$

for compatibility with the requirement

$$\int_{-\infty}^{\infty} \phi(t)dt = 1.$$

Since the $\mathbf{f}_0$ filter coefficients are fixed multiples of the $\breve{\mathbf{h}}_1$ coefficients we will also need to alter the analysis wavelet equation by a factor of 2 in order to obtain the correct orthogonality conditions (and we have done this below).

REMARK 2: We introduced the modified analysis filter coefficients $\breve{\mathbf{h}}_j(n) = \mathbf{h}_j(\ell - n)$ in order to adapt the biorthogonal filter identities to the identities needed for wavelets, but we left the synthesis filter coefficients unchanged. We could just as well have left the analysis filter coefficients unchanged and introduced modified synthesis coefficients $\breve{\mathbf{f}}_j(n) = \mathbf{f}_j(\ell - n)$.

Associated with the analyzing function $\tilde{\phi}(t)$ there must be an analyzing wavelet $\tilde{w}(t)$, with norm 1, and satisfying the *analysis wavelet equation*

$$\tilde{w}(t) = \sum_k \breve{\mathbf{h}}_1(k)\tilde{\phi}(2t-k), \tag{9.57}$$

and such that $\tilde{w}$ is orthogonal to all translations $\phi(t-k)$ of the synthesis function.

Associated with the synthesis function $\phi(t)$ there must be an synthesis wavelet $w(t)$, with norm 1, and satisfying the *synthesis wavelet equation*

$$w(t) = 2\sum_k \mathbf{f}_1(k)\phi(2t-k), \tag{9.58}$$

and such that w is orthogonal to all translations $\tilde{\phi}(t-k)$ of the analysis function.

Since $\tilde{w}(t) \perp \phi(t-m)$ for all m, the vector $\breve{\mathbf{h}}_1$ satisfies double-shift orthogonality with $\mathbf{f}_0$:

$$< \phi(t-m), \tilde{w}(t) >= \sum_k \breve{\mathbf{h}}_1(k)\mathbf{f}_0(k+2m) = 0. \tag{9.59}$$

The requirement that $\tilde{w}(t) \perp w(t-m)$ for nonzero integer m leads to double-shift orthogonality of $\breve{\mathbf{h}}_1$ to $\mathbf{f}_1$:

$$< w(t-m), \tilde{w}(t) >= \sum_k \breve{\mathbf{h}}_1(k)\mathbf{f}_1(k-2m) = \delta_{0m}. \tag{9.60}$$

Since $w(t) \perp \tilde{\phi}(t-m)$ for all m, the vector $\breve{\mathbf{h}}_0$ satisfies double-shift orthogonality with $\mathbf{f}_1$:

$$< w(t), \tilde{\phi}(t-m) >= \sum_k \breve{\mathbf{h}}_0(k)\mathbf{f}_1(k+2m) = 0. \tag{9.61}$$

The requirement that $\phi(t) \perp \tilde{\phi}(t-m)$ for nonzero integer m leads to double-shift orthogonality of $\breve{\mathbf{h}}_0$ to $\mathbf{f}_0$:

$$< \phi(t), \tilde{\phi}(t-m) >= \sum_k \breve{\mathbf{h}}_0(k)\mathbf{f}_0(k-2m) = \delta_{0m}. \tag{9.62}$$

Thus,

$$W_0 = \mathrm{Span}\{w(t-k)\}, \quad \tilde{W}_0 = \mathrm{Span}\{\tilde{w}(t-k)\}, \quad V_0 \perp \tilde{W}_0, \quad \tilde{V}_0 \perp W_0.$$

Once $w, \tilde{w}$ have been determined we can define functions

$$w_{jk}(t) = 2^{\frac{j}{2}}w(2^jt-k), \quad \tilde{w}_{jk}(t) = 2^{\frac{j}{2}}\tilde{w}(2^jt-k)$$

$$j,k = 0, \pm 1, \pm 2, \cdots,$$

and

$$\phi_{jk}(t) = 2^{\frac{j}{2}}\phi(2^j t - k), \quad \tilde{\phi}_{jk}(t) = 2^{\frac{j}{2}}\tilde{\phi}(2^j t - k)$$
$$j,k = 0, \pm 1, \pm 2, \cdots,$$

It is easy to prove the biorthogonality result.

Lemma 48

$$\begin{aligned} < w_{jk}, \tilde{w}_{j'k'} > &= \delta_{jj'}\delta_{kk'}, \qquad < \phi_{jk}, \tilde{w}_{jk'} >= 0 \\ < \phi_{jk}, \tilde{\phi}_{jk'} > &= \delta_{kk'}, \qquad < w_{jk'}, \tilde{\phi}_{jk} >= 0, \end{aligned} \tag{9.63}$$

where $j, j', k, k' = 0, \pm 1, \cdots$.

The dilation and wavelet equations extend to:

$$\tilde{\phi}_{j\ell} = 2\sum_k \check{\mathbf{h}}_0(k - 2\ell)\tilde{\phi}_{j+1,k}(t), \tag{9.64}$$

$$\phi_{j\ell} = \sum_k \mathbf{f}_0(k - 2\ell)\phi_{j+1,k}(t), \tag{9.65}$$

$$\tilde{w}_{j\ell} = \sum_k \check{\mathbf{h}}_1(k - 2\ell)\tilde{\phi}_{j+1,k}(t), \tag{9.66}$$

$$w_{j\ell} = 2\sum_k \mathbf{f}_1(k - 2\ell)\phi_{j+1,k}(t), \tag{9.67}$$

Now we have

$$V_j = \text{Span}\{\phi_{j,k}\}, \quad \tilde{V}_j = \text{Span}\{\tilde{\phi}_{j,k}\},$$
$$W_j = \text{Span}\{w_{j,k}\}, \quad \tilde{W}_j = \text{Span}\{\tilde{w}_{j,k}\}, \quad V_j \perp \tilde{W}_j, \quad \tilde{V}_j \perp W_j.$$

We can now get biorthogonal wavelet expansions for functions $f \in L^2$.

Theorem 75

$$L^2[-\infty, \infty] = V_j + \sum_{k=j}^{\infty} W_k = V_j + W_j + W_{j+1} + \cdots = \sum_{j=-\infty}^{\infty} W_j,$$

so that each $f(t) \in L^2[-\infty, \infty]$ *can be written uniquely in the form*

$$f = f_j + \sum_{k=j}^{\infty} w_k, \qquad w_k \in W_k, \; f_j \in V_j. \tag{9.68}$$

Similarly

$$L^2[-\infty,\infty]=\tilde V_j+\sum_{k=j}^{\infty}\tilde W_k=\tilde V_j+\tilde W_j+\tilde W_{j+1}+\cdots=\sum_{j=-\infty}^{\infty}\tilde W_j,$$

so that each $\tilde f(t)\in L^2[-\infty,\infty]$ *can be written uniquely in the form*

$$\tilde f=\tilde f_j+\sum_{k=j}^{\infty}\tilde w_k,\qquad \tilde w_k\in\tilde W_k,\ \tilde f_j\in\tilde V_j. \tag{9.69}$$

We have a family of new biorthogonal bases for $L^2[-\infty,\infty]$, two for each integer j:

$$\{\phi_{jk},w_{j'k'};\ \tilde\phi_{jk},\tilde w_{j'k'}:\qquad j'=j,j+1,\cdots,\quad \pm k,\pm k'=0,1,\cdots\}.$$

Let's consider the space V_j for fixed j. On the one hand we have the scaling function basis

$$\{\phi_{j,k}:\qquad \pm k=0,1,\cdots\}.$$

Then we can expand any $f_j\in V_j$ as

$$f_j=\sum_{k=-\infty}^{\infty}\tilde a_{j,k}\phi_{j,k},\quad \tilde a_{j,k}=<f_j,\tilde\phi_{j,k}> \tag{9.70}$$

On the other hand we have the wavelets basis

$$\{\phi_{j-1,k},w_{j-1,k'}:\qquad \pm k,\pm k'=0,1,\cdots\}$$

associated with the direct sum decomposition

$$V_j=W_{j-1}+V_{j-1}.$$

Using this basis we can expand any $f_j\in V_j$ as

$$f_j=\sum_{k'=-\infty}^{\infty}\tilde b_{j-1,k'}w_{j-1,k'}+\sum_{k=-\infty}^{\infty}\tilde a_{j-1,k}\phi_{j-1,k}, \tag{9.71}$$

where

$$\tilde b_{j-1,k}=<f_j,\tilde w_{j-1,k}>,\quad \tilde a_{j-1,k}=<f_j,\tilde\phi_{j-1,k}>.$$

There are exactly analogous expansions in terms of the $\tilde\phi,\tilde w$ basis.

If we substitute the relations

$$\tilde{\phi}_{j-1,\ell} = 2\sum_k \breve{\mathbf{h}}_0(k-2\ell)\tilde{\phi}_{jk}(t), \tag{9.72}$$

$$\tilde{w}_{j-1,\ell} = \sum_k \breve{\mathbf{h}}_1(k-2\ell)\tilde{\phi}_{j,k}(t), \tag{9.73}$$

into the expansion (9.70) and compare coefficients of $\phi_{j,\ell}$ with the expansion (9.71), we obtain the following fundamental recursions.

Theorem 76 *Fast Wavelet Transform.*

$$\text{Averages(lowpass)} \quad \tilde{a}_{j-1,k} = \textstyle\sum_n \breve{\mathbf{h}}_0(n-2k)\tilde{a}_{jn} \tag{9.74}$$

$$\text{Differences(highpass)} \quad \tilde{b}_{j-1,k} = \tfrac{1}{2}\textstyle\sum_n \breve{\mathbf{h}}_1(n-2k)\tilde{a}_{jn}. \tag{9.75}$$

These equations link the wavelets with the biorthogonal filter bank. Let $\mathbf{x}(k) = \tilde{a}_{jk}$ be a discrete signal. The result of passing this signal through the (time reversed) filter $\breve{\mathbf{H}}_0^T$ and then downsampling is $\mathbf{y}(k) = (\downarrow 2)\breve{\mathbf{H}}_0^T * \mathbf{x}(k) = \tilde{a}_{j-1,k}$, where $\tilde{a}_{j-1,k}$ is given by (9.74). Similarly, the result of passing the signal through the (time-reversed) filter $\breve{\mathbf{H}}_1^T$ and then downsampling is $\mathbf{z}(k) = (\downarrow 2)\breve{\mathbf{H}}_1^T * \mathbf{x}(k) = \tilde{b}_{j-1,k}$, where $\tilde{b}_{j-1,k}$ is given by (9.75).

The picture is in Figure 9.2.

We can iterate this process by inputting the output $\tilde{a}_{j-1,k}$ of the low pass filter to the filter bank again to compute $\tilde{a}_{j-2,k}, \tilde{b}_{j-2,k}$, etc. At each stage we save the wavelet coefficients $\tilde{b}_{j'k'}$ and input the scaling coefficients $\tilde{a}_{j'k'}$ for further processing, see Figure 9.3.

The output of the final stage is the set of scaling coefficients $\tilde{a}_{0k}$, assuming that we stop at $j = 0$. Thus our final output is the complete set of coefficents for the wavelet expansion

$$f_j = \sum_{j'=0}^{j-1}\sum_{k=-\infty}^{\infty} \tilde{b}_{j'k} w_{j'k} + \sum_{k=-\infty}^{\infty} \tilde{a}_{0k}\phi_{0k},$$

based on the decomposition

$$V_j = W_{j-1} + W_{j-2} + \cdots + W_1 + W_0 + V_0.$$

To derive the synthesis filter bank recursion we can substitute the inverse relation

$$\phi_{j,s} = \sum_h \left(\overline{\mathbf{f}}_0(s-2h)\phi_{j-1,h} + \overline{\mathbf{f}}_1(s-2h)w_{j-1,h}\right), \tag{9.76}$$

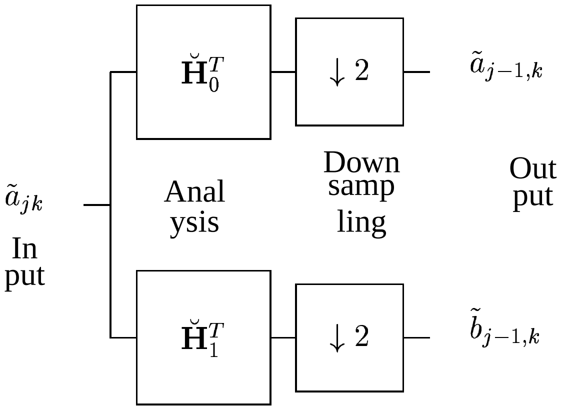

Figure 9.2: Wavelet Recursion

into the expansion (9.71) and compare coefficients of $\phi_{j-1,\ell}, w_{j-1,\ell}$ with the expansion (9.70) to obtain the inverse recursion.

Theorem 77 *Inverse Fast Wavelet Transform.*

$$\tilde{a}_{j,\ell} = \sum_k \mathbf{f}_0(\ell - 2k)\tilde{a}_{j-1,k} + \sum_k \mathbf{f}_1(\ell - 2k)\tilde{b}_{j-1,k}. \tag{9.77}$$

This is exactly the output of the synthesis filter bank shown in Figure 9.4.

Thus, for level j the full analysis and reconstruction picture is Figure 9.5.

For any $f(t) \in L^2[-\infty, \infty]$ the scaling and wavelets coefficients of f are defined by

$$\begin{aligned} \tilde{a}_{jk} &= < f, \tilde{\phi}_{jk} >= 2^{j/2}\int_{-\infty}^{\infty} f(t)\tilde{\phi}(2^j t - k)dt, \qquad (9.78) \\ \tilde{b}_{jk} &= < f, \tilde{w}_{jk}) = 2^{j/2}\int_{-\infty}^{\infty} f(t)\tilde{w}(2^j t - k)dt \end{aligned}$$

9.4.3 Sufficient Conditions for Biorthogonal Multiresolution Analysis

We have seen that the coefficients $\breve{\mathbf{h}}_i, \mathbf{f}_i$ in the analysis and synthesis dilation and wavelet equations must satisfy exactly the same double-shift orthogonality properties as those that come from biorthogonal filter banks. Now we will assume that

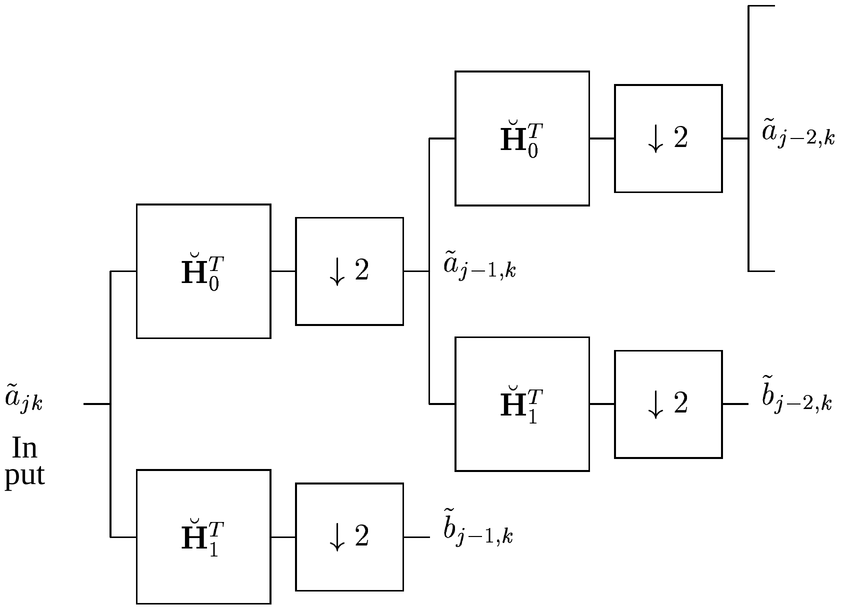

Figure 9.3: General Fast Wavelet Transform

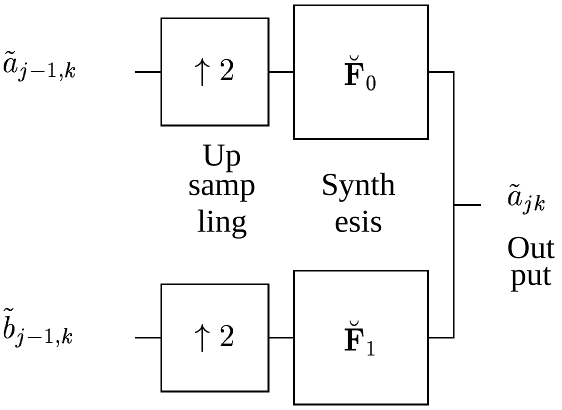

Figure 9.4: Wavelet inversion

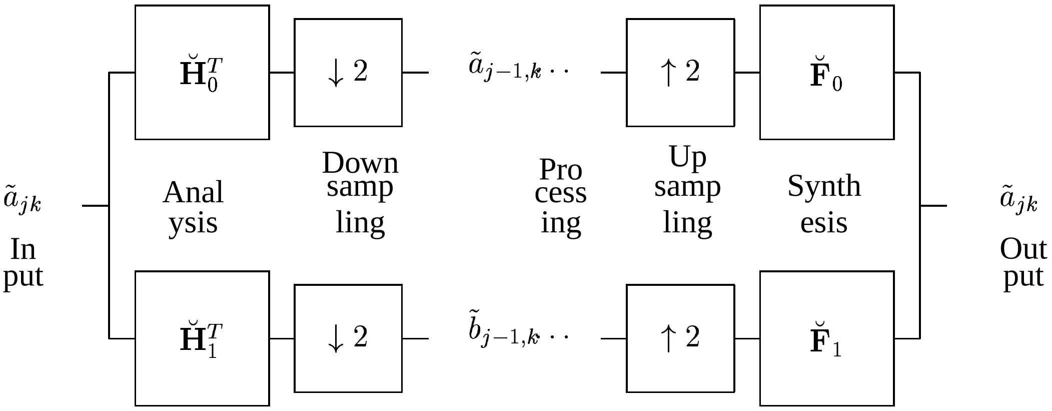

Figure 9.5: General Fast Wavelet Transform and Inversion

we have coefficients $\breve{\mathbf{h}}_i, \mathbf{f}_i$ satisfying these double-shift orthogonality properties and see if we can construct analysis and synthesis functions and wavelet functions associated with a biorthogonal multiresolution analysis.

The construction will follow from the cascade algorithm, applied to functions $\phi^{(i)}(t)$ and $\tilde{\phi}^{(i)}(t)$ in parallel. We start from the box function $\phi^{(0)}(t) = \tilde{\phi}^{(0)}(t)$ on $[0, 1]$. We apply the low pass filter $\mathbf{F}_0$, recursively and with scaling, to the $\phi^{(i)}(t)$, and the low pass filter $\breve{\mathbf{H}}_0$, recursively and with scaling, to the $\tilde{\phi}^{(i)}(t)$. (For each pass of the algorithm we rescale the iterate by multiplying it by $\frac{1}{\sqrt{2}}$ to preserve normalization.)

Theorem 78 *If the cascade algorithm converges uniformly in L^2 for both the analysis and synthesis functions, then the limit functions $\phi(t), \tilde{\phi}$ and associated wavelets $w(t), \tilde{w}(t)$ satisfy the orthogonality relations*

$$< w_{jk}, \tilde{w}_{j'k'} >= \delta_{jj'}\delta_{kk'}, \qquad < \phi_{jk}, \tilde{\phi}_{jk'} >= \delta_{kk'},$$

$$< \phi_{jk}, \tilde{w}_{j,k'} >= 0, \quad < w_{jk}, \tilde{\phi}_{j,k'} >= 0$$

where $j, j', \pm k, \pm k' = 0$.

PROOF: There are only three sets of identities to prove:

$$1. \qquad \int_{-\infty}^{\infty} \phi(t-n)\tilde{\phi}(t-m)dt = \delta_{nm}$$

$$2. \qquad \int_{-\infty}^{\infty} \phi(t-n)\tilde{w}(t-m)dt = 0$$

$$3. \qquad \int_{-\infty}^{\infty} w(t-n)\tilde{w}(t-m)dt = \delta_{nm}.$$

The rest are duals of these, or immediate.

1. We will use induction. If 1. is true for the functions $\phi^{(i)}(t), \tilde{\phi}^{(i)}(t)$ we will show that it is true for the functions $\phi^{(i+1)}(t), \tilde{\phi}^{(i+1)}(t)$. Clearly it is true for $\phi^{(0)}(t), \tilde{\phi}^{(0)}(t)$. Now

$$\int_{-\infty}^{\infty} \phi^{(i+1)}(t-n)\tilde{\phi}^{(i+1)}(t-m)dt =< \phi^{(i+1)}_{0n}, \tilde{\phi}^{(i+1)}_{0m} >$$

$$=< \sum_k \mathbf{f}_0(k)\phi^{(i)}_{1,2n+k}, \sum_\ell \breve{\mathbf{h}}_0(\ell)\tilde{\phi}^{(i)}_{1,2m+\ell} >$$

$$= \sum_{k\ell} \mathbf{f}_0(k)\mathbf{h}_0\breve{(}\ell) < \phi^{(i)}_{1,2n+k}, \tilde{\phi}^{(i)}_{1,2m+\ell} >= \sum_k \mathbf{f}_0(k)\breve{\mathbf{h}}_0(k-2(m-n)) = \delta_{nm}.$$

Since the convergence is L^2, these orthogonality relations are also valid in the limit for $\phi(t), \tilde{\phi}(t)$.

2.
$$\int_{-\infty}^{\infty}\phi^{(i+1)}(t-n)\tilde{w}^{(i+1)}(t-m)dt =< \phi^{(i+1)}_{0n}, \tilde{w}^{(i+1)}_{0m} >$$
$$=< \sum_k \mathbf{f}_0(k)\phi^{(i)}_{1,2n+k}, \sum_\ell \breve{\mathbf{h}}_1(\ell)\tilde{w}^{(i)}_{1,2m+\ell} >$$
$$= \sum_{k\ell} \mathbf{f}_0(k)\breve{\mathbf{h}}_1(\ell) < \phi^{(i)}_{1,2n+k}, \tilde{w}^{(i)}_{1,2m+\ell} >= 2\sum_k \mathbf{f}_0(k)\breve{\mathbf{h}}_1(k-2(m-n)) = 0,$$

because of the double-shift orthogonality of $\mathbf{f}_0$ and $\breve{\mathbf{h}}_1$.

3.
$$\int_{-\infty}^{\infty} w^{(i+1)}(t-n)\tilde{w}^{(i+1)}(t-m)dt =< w^{(i+1)}_{0n}, \tilde{w}^{(i+1)}_{0m} >$$
$$= \sum_k \mathbf{f}_1(k)\breve{\mathbf{h}}_1(k-2(m-n)) = \delta_{nm},$$

because of the double-shift orthonormality of $\mathbf{f}_1$ and $\breve{\mathbf{h}}_1$.

Q.E.D.

Theorem 79 *Suppose the filter coefficients satisfy double-shift orthogonality conditions (9.48), (9.49), (9.50) and (9.51), as well as the conditions of Theorem 55 for the matrices* $\mathbf{T} = (\downarrow 2)2\mathbf{F}_0\overline{\mathbf{F}}_0^{\rm tr}$ *and* $\tilde{\mathbf{T}} = (\downarrow 2)2\breve{\mathbf{H}}_0\overline{\breve{\mathbf{H}}}_0^{\rm tr}$, *which guarantee* L^2 *convergence of the cascade algorithm. Then the synthesis functions* $\phi(t-k), w(t-k)$ *are orthogonal to the analysis functions* $\tilde{\phi}(t-\ell), \tilde{w}(t-\ell)$, *and each scaling space is orthogonal to the dual wavelet space:*

$$V_j \perp \tilde{W}_j, \qquad W_j \perp \tilde{V}_j. \tag{9.79}$$

Also
$$V_j + W_j = V_{j+1}, \qquad \tilde{V}_j + \tilde{W}_j = \tilde{V}_{j+1}$$
where the direct sums are, in general, not orthogonal.

Corollary 23 *The wavelets*

$$w_{jk} = 2^{\frac{j}{2}}w(2^j - k), \qquad \tilde{w}_{jk} = 2^{\frac{j}{2}}\tilde{w}(2^j - k),$$

are biorthogonal bases for L^2*:*

$$\int_{-\infty}^{\infty} w_{jk}(t)\tilde{w}_{j'k'}(t)dt = \delta_{jj'}\delta_{kk'}. \tag{9.80}$$

A TOY (BUT VERY EXPLICIT) EXAMPLE: We will construct this example by first showing that it is possible to associate a scaling function $\Phi(t)$ the the identity low pass filter $H_0(\omega) = H_0(z) \equiv 1$. Of course, this really isn't a low pass filter, since $H_0(\pi) \neq 0$ in the frequency domain and the scaling function will not be a function at all, but a distribution or "generalized function". If we apply the cascade algorithm construction to the identity filter $H_0(\omega) \equiv 1$ in the frequency domain, we easily obtain the Fourier transform of the scaling function as $\hat{\phi}(\omega) \equiv 1$. Since $\hat{\phi}(\omega)$ isn't square integrable, there is no true function $\phi(t)$. However there is a distribution $\phi(t) = \delta(t)$ with this transform. Distributions are linear functionals on function classes, and are informally defined by the values of integrals of the distribution with members of the function class.

Recall that $f \in L^2[-\infty, \infty]$ belongs to the *Schwartz class* if f is infinitely differentiable everywhere, and there exist constants $C_{n,q}$ (depending on f) such that $|t^n \frac{d^q}{dt^q} f| \leq C_{n,q}$ on R for each $n, q = 0, 1, 2, \cdots$. One of the pleasing features of this space of functions is that f belongs to the class if and only if $\hat{f}$ belongs to the class. We will define the distribution $\phi(t)$ by its action as a linear functional on the Schwartz class. Consider the Parseval formula

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

where f belongs to the Schwartz class. We will *define* the integral on the left-hand side of this expression by the integral on the right-hand side. Thus

$$\int_{-\infty}^{\infty} \phi(t) f(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\phi}(\omega) \hat{f}(\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) d\omega = f(0)$$

from the inverse Fourier transform. This functional is the *Dirac Delta Function*, $\phi(t) = \delta(t)$. It picks out the value of the integrand at $t = 0$.

We can use the standard change of variables formulas for integrals to see how distributions transform under variable change. Since $H_0(\omega) \equiv 1$ the corresponding filter has $\mathbf{h}_0(0) = 1$ as its only nonzero coefficient. Thus the dilation equation in the signal domain is $\phi(t) = 2\phi(2t)$. Let's show that $\phi(t0 = \delta(t)$ is a solution of this equation. On one hand

$$\int_{-\infty}^{\infty} \phi(t) f(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\phi}(\omega) = f(0)$$

for any function $f(t)$ in the Schwartz class. On the other hand (for $\tau = 2t$)

$$\int_{-\infty}^{\infty} 2\phi(2t) f(t) dt = \int_{-\infty}^{\infty} \phi(\tau) f(\frac{\tau}{2}) d\tau = g(0) = f(0)$$

since $g(\tau) = f(\frac{\tau}{2})$ belongs to the Schwartz class. The the distributions $\delta(t)$ and $2\delta(2t)$ are the same.

Now we proceed with our example and consider the biorthogonal scaling functions and wavelets determined by the biorthogonal filters

$$H_0(z) = 1, \qquad F_0(z) = \frac{1}{2} + z^{-1} + \frac{1}{2}z^{-2}$$

Then alias cancellation gives $H_1(z) = F_0(-z), F_1(z) = -H_0(-z)$, so

$$H_1(z) = \frac{1}{2} - z^{-1} + \frac{1}{2}z^{-2}, \qquad F_1(z) = -1.$$

The low pass analysis and synthesis filters are related to the half band filter P_0 by

$$P_0(z) = H_0(z)F_0(z), \qquad \text{where } P_0(1) = 2, \quad P_0(z) - P_0(-z) = 2z^{-\ell}$$

and ℓ is the delay. Then alias cancellation gives $H_1(z) = F_0(-z), F_1(z) = -H_0(-z)$, so

$$H_1(z) = \frac{1}{2} - z^{-1} + \frac{1}{2}z^{-2}, \qquad F_1(z) = -1,$$

and we find that $P_0(z) = \frac{1}{2} + z^{-1} + \frac{1}{2}z^{-2}$, so $\ell = 1$.

To pass from the biorthogonal filters to coefficient identities needed for the construction of wavelets we have to modify the filter coefficients. In the notes above I modified the analysis coefficients to obtain new coefficients $\breve{\mathbf{h}}_j(n) = \mathbf{h}_j(\ell - n)$, $j = 0, 1$, and left the synthesis coefficients as they are. Since the choice of what is an analysis filter and what is a synthesis filter is arbitrary, I could just as well have modified the synthesis coefficients to obtain new coefficients $\breve{\mathbf{f}}_j(n) = \mathbf{f}_j(\ell - n)$, $j = 0, 1$, and left the analysis coefficients unchanged. In this problem it is important that one of the low pass filters be $H_0(z) = 1$ (so that the delta function is the scaling function). If we want to call that an analysis filter, then we have to modify the synthesis coefficients.

Thus the nonzero analysis coefficients are

$$(\mathbf{h}_0(0)) = (1) \qquad \text{and } (\mathbf{h}_1(0), \mathbf{h}_1(1), \mathbf{h}_1(2)) = (\frac{1}{2}, -1, \frac{1}{2}).$$

The nonzero synthesis coefficients are

$$(\mathbf{f}_1(0)) = (-1) \qquad \text{and } (\mathbf{f}_0(0), \mathbf{f}_0(1), \mathbf{f}_0(2)) = (\frac{1}{2}, 1, \frac{1}{2}).$$

Since $\breve{\mathbf{f}}_j(n) = \mathbf{f}_j(\ell - n) = \mathbf{f}_j(1-n)$ for $j = 0, 1$ we have nonzero components $(\breve{\mathbf{f}}_0(-1), \breve{\mathbf{f}}_0(0), \breve{\mathbf{f}}_0(1)) = (\frac{1}{2}, 1, \frac{1}{2})$ and $(\breve{\mathbf{f}}_1(1)) = (-1)$, so the modified synthesis filters are $\breve{F}_0(z) = \frac{1}{2}z + 1 + \frac{1}{2}z^{-1}$, and $\breve{F}_1(z) = -z^{-1}$.

The analysis dilation equation is $\tilde{\phi}(t) = 2\tilde{\phi}(2t)$ with solution $\tilde{\phi}(t) = \delta(t)$, the Dirac delta function.

The synthesis dilation equation should be

$$\phi(t) = \sum_k \breve{\mathbf{f}}_0(k)\phi(2t-k),$$

or

$$\phi(t) = \frac{1}{2}\phi(2t+1) + \phi(2t) + \frac{1}{2}\phi(2t-1).$$

It is straightforward to show that the hat function (centered at $t = 0$)

$$\phi(t) = \begin{cases} 1+t & -1 \le t < 0 \\ 1-t & 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases}$$

is the proper solution to this equation.

The analysis wavelet equation is

$$\tilde{w}(t) = \sum_k \mathbf{h}_1(k)\tilde{\phi}(2t-k),$$

or

$$\tilde{w}(t) = \frac{1}{2}\tilde{\phi}(2t) - \tilde{\phi}(2t-1) + \frac{1}{2}\tilde{\phi}(2t-2)$$

The synthesis wavelet equation is

$$w(t) = 2\sum_k \breve{\mathbf{f}}_1(k)\phi(2t-k). \qquad \text{or } w(t) = -2\phi(2t-1).$$

Now it is easy to verify explicitly the biorthogonality conditions

$$\int \phi(t)\tilde{\phi}(t-k)dt = \int w(t)\tilde{w}(t-k)dt = \delta(k),$$

$$\int \phi(t)\tilde{w}(t-k)dt = \int \tilde{\phi}(t)w(t-k)dt = 0.$$

9.4.4 Splines

A *spline* $f(t)$ of order N on the grid of integers is a piecewise polynomial

$$f_j(t) = a_{0,j} + a_{1,j}(t-j) + \cdots a_{N,j}(t-j)^N, \qquad j \le t \le j+1,$$

such that the pieces fit together smoothly at the gridpoints:

$$f_j(j) = f_{j-1}(j), f_j'(j) = f_{j-1}'(j), \cdots, f_j^{(N-1)}(j) = f_{j-1}^{(N-1)}(j), \qquad j = 0, \pm 1, \cdots.$$

Thus $f(t)$ has $N-1$ continuous derivatives for all t. The Nth derivative exists for all noninteger t and for $t = j$ the right and left hand derivatives $f^{(N)}(j+0), f^{(N)}(j-0)$ exist. Furthermore, we assume that a spline has compact support. Splines are widely used for approximation of functions by interpolation. That is, if $F(t)$ is a function taking values $F(j)$ at the gridpoints, one approximates F by an N-spline f that takes the same values at the gridpoints: $f(j) = F(j)$, for all j. Then by subdividing the grid (but keeping N fixed) one can show that these N-spline approximations get better and better for sufficiently smooth F. The most commonly used splines are the *cubic splines*, where $N = 3$.

Splines have a close relationship with wavelet theory. Usually the wavelets are biorthogonal, rather than orthogonal, and one set of N-splines can be associated with several sets of biorthogonal wavelets. We will look at a few of these connections as they relate to multiresolution analysis. We take our low pass space V_0 to consist of the $N-1$-splines on unit intervals and with compact support. The space V_1 will then contain the $N-1$-splines on half-intervals, etc. We will find a basis $\phi(t-k)$ for V_0 (but usually not an ON basis).

We have already seen examples of splines for the simplest cases. The 0-splines are piecewise constant on unit intervals. This is just the case of Haar wavelets. The scaling function $\phi_0(t)$ is just the box function

$$\phi_0(t) = B(t) = \begin{cases} 1, & 0 \le t < 1 \\ 0, & \text{otherwise} \end{cases}$$

Here, $\hat{\phi}_0(\omega) = \hat{B}(t) = \frac{1}{i\omega}(1 - e^{-i\omega})$, essentially the sinc function. Here the set $\phi(t-k)$ is an ON basis for V_0.

The 1-spines are continuous piecewise linear functions. The functions $f(t) \in V_0$ are determined by their values $f(k)$ at the integer points, and are linear between each pair of values:

$$f(t) = [f(k+1) - f(k)](t-k) + f(k) \quad \text{for } k \le t \le k+1.$$

The scaling function ϕ_1 is the *hat function*. The hat function $H(t)$ is the continuous piecewise linear function whose values on the integers are $H(k) = \delta_{1k}$, i.e., $H(1) = 1$ and $H(t)$ is zero on the other integers. The support of $H(t)$ is the open interval $-0 < t < 2$. Furthermore, $H(t) = \phi_1(t) = (B * B)(t)$, i.e., the hat function is the convolution of two box functions. Moreover, $\hat{\phi}_1(\omega) = \hat{B}^2(t) = (\frac{1}{i\omega})^2(1 - e^{-i\omega})^2$ Note that if $f \in V_0$ then we can write it uniquely in the form

$$f(t) = \sum_k f(k)H(t - k + 1).$$

All multiresolution analysis conditions are satisfied, except for the ON basis requirement. The integer translates of the hat function do define a Riesz basis for V_0 (though we haven't completely proved it yet) but it isn't ON because the inner product $(H(t), H(t-1)) \neq 0$. A scaling function does exist whose integer translates form an ON basis, but its support isn't compact. It is usually simpler to stick with the nonorthogonal basis, but embed it into a biorthogonal multiresolution structure, as we shall see.

Based on our two examples, it is reasonable to guess that an appropriate scaling function for the space V_0 of $N-1$-splines is

$$\phi_{N-1}(t) = (B*B*\cdots*B)(t) \quad \text{N times}, \qquad \hat{\phi}_{N-1}(\omega) = \hat{B}^N(t) = (\frac{1}{i\omega})^N(1-e^{-i\omega})^N. \tag{9.81}$$

This special spline is called a *B-spline*, where the B stands for *basis*.

Let's study the properties of the B-spline. First recall the definition of the convolution:

$$f * g(t) = \int_{-\infty}^{\infty} f(t-x)g(x)dx = \int_{-\infty}^{\infty} f(x)g(t-x)dx.$$

If $f = B$, the box function, then

$$B * g(t) = \int_{t-1}^{t} g(x)dx = \int_0^1 g(t-x)dx.$$

Now note that $\phi_N(t) = B * \phi_{N-1}(t)$, so

$$\phi_N(t) = \int_{t-1}^{t} \phi_{N-1}(x)dx. \tag{9.82}$$

Using the fundamental theorem of calculus and differentiating, we find

$$\phi'_N(t) = \phi_{N-1}(t) - \phi_{N-1}(t-1). \tag{9.83}$$

Now $\phi_0(t)$ is piecewise constant, has support in the interval $[0, 1)$ and discontinuities 1 at $t = 0$ and -1 at $t = 1$.

Theorem 80 *The function $\phi_N(t)$ has the following properties:*

1. *It is a spline, i.e., it is piecewise polynomial of order N.*
2. *The support of $\phi_N(t)$ is contained in the interval $[0, N+1)$.*
3. *The jumps in the Nth derivative at $t = 0, 1, \cdots, N+1$ are the alternating binomial coefficients $(-1)^t \left(\begin{array}{c} N \\ t \end{array} \right)$.*

PROOF: By induction on N. We observe that the theorem is true for $N = 0$. Assume that it holds for $N = M - 1$. Since $\phi_{M-1}(t)$ is piecewise polynomial of order $M - 1$ and with support in $[0, M)$, it follows from (9.82) that $\phi_M(t)$ is piecewise polynomial of order M and with support in $[0, M+1)$. Denote by $[\phi_N^{(M)}(j)] = \phi_N^{(M)}(j+0) - \phi_N^{(M)}(j-0)$ the jump in $\phi_N^{(M)}(t)$ at $t = j$. Differentiating (9.83) $M - 1$ times for $N = M$, we find

$$[\phi_N^{(M)}(j)] = [\phi_N^{(M-1)}(j)] - [\phi_N^{(M-1)}(j-1)]$$

for $j = 0, 1, \cdots M + 1$ where

$$[\phi_N^{(M)}(j)] = (-1)^j \left(\begin{array}{c} M-1 \\ j \end{array} \right) - (-1)^{j-1} \left(\begin{array}{c} M-1 \\ j-1 \end{array} \right)$$

$$= (-1)^j \left(\frac{(M-1)!}{j!(M-j-1)!} + \frac{(M-1)!}{(j-1)!(M-j)!} \right)$$

$$= \frac{(-1)^j M!}{j!(M-j)!} \left(\frac{M-j}{M} + \frac{j}{M} \right) = (-1)^j \left(\begin{array}{c} M \\ j \end{array} \right).$$

Q.E.D.

Normally we start with a low pass filter $\mathbf{H}$ with finite impulse response vector $\mathbf{h}(n)$ and determine the scaling function via the dilation equation and the cascade algorithm. Here we have a different problem. We have a candidate B-spline scaling function $\phi_{N-1}(t)$ with Fourier transform

$$\hat{\phi}_{N-1}(\omega) = (\frac{1}{i\omega})^N (1 - e^{-i\omega})^N.$$

What is the associated low pass filter function $H(\omega)$? The dilation equation in the Fourier domain is

$$\hat{\phi}_{N-1}(\omega) = H(\frac{\omega}{2})\hat{\phi}_{N-1}(\frac{\omega}{2}).$$

Thus

$$(\frac{1}{i\omega})^N(1-e^{-i\omega})^N = H(\frac{\omega}{2})(\frac{1}{i\omega/2})^N(1-e^{-i\omega/2})^N.$$

Solving for $H(\omega/2)$ and rescaling, we find

$$H(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^N$$

or

$$H(z) = \left(\frac{1+z^{-1}}{2}\right)^N.$$

This is as nice a low pass filter as you could ever expect. *All* of its zeros are at -1! We see that the finite impulse response vector $\mathbf{h}(n) = \frac{1}{2^N}\left(\begin{array}{c} N \\ n \end{array}\right)$, so that the dilation equation for a spline is

$$\phi_{N-1}(t) = 2^{1-N}\sum_{k=0}^{N}\left(\begin{array}{c} N \\ n \end{array}\right)\phi_{N-1}(2t-k). \qquad (9.84)$$

For convenience we list some additional properties of B-splines. The first two properties follow directly from the fact that they are scaling functions, but we give direct proofs anyway.

Lemma 49 *The B-spline $\phi_N(t)$ has the properties:*

1. $\int_{-\infty}^{\infty}\phi_N(t)dt = 1$.

2. $\sum_k \phi_N(t+k) = 1$.

3. $\phi_N(t) = \phi_N(N+1-t)$, *for* $N = 1, 2, \cdots$.

4. *Let*

$$\chi_k(t) = \begin{cases} 1, & t \geq k \\ 0, & \text{otherwise.} \end{cases}$$

$\phi_N(t) = \sum_{k=0}^{N+1}\left(\begin{array}{c} N+1 \\ k \end{array}\right)(-1)^k\frac{(t-k)^N}{N!}\chi_k(t)$ *for* $N = 1, 2, \cdots$.

5. $\phi_N(t) \geq 0$.

6. $\mathbf{a}(k) = \int_{-\infty}^{\infty}\phi_N(t)\phi_N(t+k)dt = \phi_{2N+1}(N+k+1)$.

PROOF:

1. $\int_{-\infty}^{\infty}\phi_N(t)dt = \hat{\phi}_N(0) = 1$.

2. $$\sum_k \phi_N(t+k) = \sum_k \int_{t+k-1}^{t+k} \phi_{N-1}(x)dx = \int_{-\infty}^{\infty} \phi_{N-1}(t)dt = 1.$$

3. Use induction on N. The statement is obviously true for $N = 1$. Assume it holds for $N = M - 1$. Then
$$\phi_M(M+1-t) = \int_{M-t}^{M+1-t} \phi_{M-1}(x)dx = \int_{t-1}^{t} \phi_{M-1}(M-u)du$$
$$= \int_{t-1}^{t} \phi_{M-1}(u)du = \phi_M(t)$$
.

4. From the 3rd property in the preceding theorem, we have
$$\phi_N^{(N)}(t) = \sum_{k=0}^{N+1} \begin{pmatrix} N+1 \\ k \end{pmatrix} (-1)^k \chi_k(t).$$
Integrating this equation with respect to t, from $-\infty$ to t, N times, we obtain the desired result.

5. Follows easily, by induction on N.

6. The Fourier transform of $\phi_N(t)$ is $\hat{\phi}_N(\omega) = \hat{\phi}_0^{N+1}(\omega) = (\frac{1}{2\pi i\omega})^{N+1}(1 - e^{-i\omega})^{N+1}$ and the Fourier transform of $\phi_N(t+k)$ is $e^{ik\omega}\hat{\phi}_N(\omega)$. Thus the Plancherel formula gives
$$\int_{-\infty}^{\infty} \phi_N(t)\phi_N(t+k)dt = 2\pi \int_{-\infty}^{\infty} |\hat{\phi}_N(\omega)|^2 e^{ik\omega} d\omega =$$
$$2\pi \int_{-\infty}^{\infty} (\frac{1}{2\pi i\omega})^{2N+2}(1 - e^{-i\omega})^{2N+2} e^{i(N+k+1)\omega} d\omega = \phi_{2N+1}(N+k+1).$$

Q.E.D.

An N-spline $f(t)$ in V_0 is uniquely determined by the values $f(j)$ it takes at the integer gridpoints. (Similarly, functions $f(t)$ in V_J are uniquely determined by the values $f(k/2^J)$ for integer k.)

Lemma 50 *Let $f, \tilde{f} \in V_0$ be N-splines such that $f(j) = \tilde{f}(j)$ for all integers j. Then $f(t) \equiv \tilde{f}(t)$.*

PROOF: Let $g(t) = f(t) - \tilde{f}(t)$. Then $g \in V_0$ and $g(j) = 0$ for all integers j. Since $g \in V_0$, it has compact support. If g is not identically 0, then there is a least integer J such that $g_J \not\equiv 0$. Here

$$g_J(t) = a_{0,j} + a_{1,j}(t-J) + \cdots + a_{N,J}(t-J)^N, \qquad J \le t \le J+1.$$

However, since $g_{J-1}(t) \equiv 0$ and since

$$g_J(J) = g_{J-1}(J), g'_J(J) = g'_{J-1}(J), \cdots, g_J^{(N-1)}(J) = g_{J-1}^{(N-1)}(J),$$

we have $g_J(t) = a_{N,J}(t-J)^N$. However $g_J(J+1) = 0$, so $a_{N,J} = 0$ and $g_J(t) \equiv 0$, a contradiction. Thus $g(t) \equiv 0$. Q.E.D.

Now let's focus on the case $N = 3$. For cubic splines the finite impulse response vector is $\mathbf{h}(n) = \frac{1}{16}(1, 4, 6, 4, 1)$ whereas the cubic B-spline $\phi_3(t)$ has jumps $1, -4, 6, -4, 1$ in its 3rd derivative at the gridpoints $t = 0, 1, 2, 3, 4$, respectively. The support of $\phi_3(t)$ is contained in the interval $[0, 4)$, and from the second lemma above it follows easily that $\phi_3(0) = \phi_3(4) = 0$, $\phi_3(1) = \phi_3(3) = 1/6$. Since the sum of the values at the integer gridpoints is 1, we must have $\phi_3(2) = 4/6$. We can verify these values directly from the fourth property of the preceding lemma.

We will show directly, i.e., without using wavelet theory, that the integer translates $\phi_3(t-k)$ form a basis for the resolution space V_0 in the case $N = 3$. This means that any $f(t) \in V_0$ can be expanded uniquely in the form $f(t) = \sum_k s(k)\phi_3(t-k)$ for expansion coefficients $s(k)$ and all t. Note that for fixed t, at most 4 terms on the right-hand side of this expression are nonzero. According to the previous lemma, if the right-hand sum agrees with $f(t)$ at the integer gridpoints then it agrees everywhere. Thus it is sufficient to show that given the input $\mathbf{f}(j) = f(j)$ we can always solve the equation

$$f(j) = \sum_k s(k)\phi_3(j-k), \qquad j = 0, \pm 1, \cdots \tag{9.85}$$

for the vector $\mathbf{s}(k) = s(k)$, $k = 0, \pm 1, \cdots$. We can write (9.85) as a convolution equation $\mathbf{f} = \mathbf{b} * \mathbf{s}$ where

$$\mathbf{b} = (\mathbf{b}(0), \mathbf{b}(1), \mathbf{b}(2), \mathbf{b}(3), \mathbf{b}(4)) = (0, \frac{1}{6}, \frac{4}{6}, \frac{1}{6}, 0).$$

We need to invert this equation and solve for $\mathbf{s}$ in terms of $\mathbf{f}$. Let B be the FIR filter with impulse response vector $\mathbf{b}$. Passing to the frequency domain, we see that (9.85) takes the form

$$F(\omega) = B(\omega)S(\omega), \quad F(\omega) = \sum_j \mathbf{f}(j)e^{-ij\omega}, \quad S(\omega) = \sum_k \mathbf{s}(k)e^{-ik\omega},$$

and

$$B(\omega) = \sum_j \mathbf{b}(j)e^{-ij\omega} = \frac{e^{-i\omega}}{6}(1 + 4e^{-i\omega} + e^{-2i\omega}) = \frac{e^{-2i\omega}}{3}(2 + \cos\omega).$$

Note that $B(\omega)$ is bounded away from zero for all ω, hence B is invertible and has a bounded inverse B^{-1} with

$$B^{-1}(\omega) = \frac{3e^{2i\omega}}{2+\cos\omega} = \frac{6e^{2i\omega}}{(2+\sqrt{3})(1+[2-\sqrt{3}]e^{i\omega})(1+[2-\sqrt{3}]e^{-i\omega})}$$

$$= -\sqrt{3}e^{2i\omega}\left(\frac{\gamma e^{i\omega}}{1+[2-\sqrt{3}]e^{i\omega}} - \frac{1}{1+[2-\sqrt{3}]e^{-i\omega}}\right) = \sum_n \mathbf{c}(n)e^{in\omega}$$

where

$$\mathbf{c}(n) = \begin{cases} \sqrt{3}(2+\sqrt{3})^{2-n}(-1)^n & \text{if } n \geq 3 \\ \sqrt{3}(2+\sqrt{3})^{n-2}(-1)^n & \text{if } n \leq 2. \end{cases}$$

Thus,

$$\mathbf{s} = \mathbf{c} * \mathbf{f} \quad \text{or} \quad \mathbf{s}(k) = \sum_j \mathbf{c}(k-j)\mathbf{f}(j).$$

Note that B^{-1} is an infinite impulse response (IIR) filter.

The integer translates $\phi_N(t-k)$ of the B-splines form a Riesz basis of V_0 for each N. To see this we note from Section 8.1 that it is sufficient to show that the infinite matrix of inner products

$$A_{ij} = \int_{-\infty}^{\infty} \phi_N(t-i)\phi_N(t-j)dt = \int_{-\infty}^{\infty} \phi_N(t)\phi_N(t+i-j)dt = \mathbf{a}(i-j)$$

has positive eigenvalues, bounded away from zero. We have studied matrices of this type many times. Here, A is a Toeplitz matrix with associated impulse vector $\mathbf{a}(k)$ The action of A on a column vector $\mathbf{x}$, $\mathbf{y} = A\mathbf{x}$, is given by the convolution $\mathbf{y}(n) = \mathbf{a} * \mathbf{x}(n)$. In frequency space the action of A is given by multiplication by the function

$$A(\omega) = \sum_k \mathbf{a}(k)e^{-ik\omega} = \sum_{n=-\infty}^{\infty} |\hat{\phi}_N(\omega + 2\pi n)|^2.$$

Now

$$|\hat{\phi}_N(\omega + 2\pi n)|^2 = \left(\frac{4\sin^2(\frac{\omega}{2})}{(\omega + 2\pi n)^2} \right)^{N+1} . \qquad (9.86)$$

On the other hand, from (49) we have

$$A(\omega) = \sum_k \phi_{2N+1}(N{+}k{+}1)e^{-ik\omega} = \phi_{2N+1}(N{+}1){+}2\sum_{k>0}\phi_{2N+1}(N{+}k{+}1)\cos(k\omega).$$

Since all of the inner products are nonnegative, and their sum is 1, the maximum value of $A(\omega)$, hence the norm of A, is $A(0) = B = 1$. It is now evident from (9.86) that $A(\omega)$ is bounded away from zero for every N.(Indeed the term (9.86) alone, with $n = 0$, is bounded away from zero on the interval $[-\pi, \pi]$.) Hence the translates always form a Riesz basis.

We can say more. Computing $A'(\omega)$ by differentiating term-by-term, we find

$$A'(\omega) = -2\sum_{k>0}\phi_{2N+1}(N + k + 1)k\sin(k\omega).$$

Thus, $A(\omega)$ has a critical point at $\omega = 0$ and at $\omega = \pi$. Clearly, there is an absolute maximum at $\omega = 0$. It can be shown that there is an absolute minimum at $\omega = \pi$. Thus $A_{\text{min}} = \sum_k \phi_{2N+1}(N + k + 1)(-1)^k$.

Although the B-spline scaling function integer translates don't form an ON basis, they (along with any other Riesz basis of translates) can be "orthogonalized" by a simple construction in the frequency domain. Recall that the necessary and sufficient condition for the translates $\phi(t - k)$ of a scaling function to be an ON basis for V_0 is that

$$A(\omega) = \sum_{n=-\infty}^{\infty} |\hat{\phi}(\omega + 2\pi n)|^2 \equiv 1.$$

In general this doesn't hold for a Riesz basis. However, for a Riesz basis, we have $|A(\omega)| > 0$ for all ω. Indeed, in the B-spline case we have that $|A(\omega)|$ is bounded away from zero. Thus, we can define a modified scaling function $\tilde{\phi}_N(t)$ by

$$\hat{\tilde{\phi}}_N(\omega) = \frac{\hat{\phi}_N(\omega)}{\sqrt{|A(\omega)|}} = \frac{\hat{\phi}_N(\omega)}{\sqrt{\sum_{n=-\infty}^{\infty} |\hat{\phi}(\omega + 2\pi n)|^2}}$$

so that $\tilde{\phi}_N(t)$ is square integrable and $\tilde{A}(\omega) \equiv 1$. If we carry this out for the B-splines we get ON scaling functions $\tilde{\phi}_N(t - k)$ and wavelets, but with infinite support in the time domain. Indeed, from the explicit formulas that we have derived for $A(\omega)$ we can expand $1/\sqrt{|A(\omega)|}$ in a Fourier series

$$\frac{1}{\sqrt{|A(\omega)|}} = \sum_{k=-\infty}^{\infty} e_k e^{ik\omega}$$

so that

$$\hat{\tilde{\phi}}_N(\omega) = \sum_{k=-\infty}^{\infty} e_k e^{ik\omega} \hat{\phi}_N(\omega),$$

or

$$\tilde{\phi}_N(t) = \sum_{k=-\infty}^{\infty} e_k \phi_N(t-k).$$

This expresses the scaling function generating an ON basis as a convolution of translates of the B-spline. Of course, the new scaling function does not have compact support.

Usually however, the B-spline scaling function $\phi_N(t)$ is embedded in a family of biorthogonal wavelets. There is no unique way to do this. A natural choice is to have the B-spline as the scaling function associated with the synthesis filter F_0 Since $P_0(z) = H_0(z)F_0(z)$, the half band filter P_0 must admit $\left(\frac{1+z^{-1}}{2}\right)^{N+1}$ as a factor, to produce the B-spline. If we take the half band filter to be one from the Daubechies (maxflat) class then the factor $H_0(z)$ must be of the form $\left(\frac{1+z^{-1}}{2}\right)^{2p-N-1} Q_{2p-2}(z)$. We must then have $2p > N+1$ so that H_0 will have a zero at $z = -1$. The smallest choice of p may not be appropriate, because Condition E for a stable basis and convergence of the cascade algorithm, and the Riesz basis condition for the integer translates of the analysis scaling function may not be satisfied. For the cubic B-spline, a choice that works is

$$F_0(z) = \left(\frac{1+z^{-1}}{2}\right)^4, \qquad H_0(z) = \left(\frac{1+z^{-1}}{2}\right)^4 Q_6(z),$$

i.e., $p = 4$. This 11/5 filter bank corresponds to Daubechies D_8 (which would be 8/8). The analysis scaling function is *not* a spline.

9.5 Generalizations of Filter Banks and Wavelets

In this section we take a brief look at some extensions of the theory of filter banks and of wavelets. In one case we replace the integer 2 by the integer M, and in the other we extend scalars to vectors. These are, most definitely, topics of current research.

9.5.1 M Channel Filter Banks and M Band Wavelets

Although 2 channel filter banks are the norm, M channel filter banks with $M > 2$ are common. There are M analysis filters and the output from each is downsampled $(\downarrow M)$ to retain only $1/M$ th the information. For perfect reconstruction, the downsampled output from each of the M analysis filters is upsampled $(\uparrow M)$, passed through a synthesis filter, and then the outputs from the M synthesis filters are added to produce the original signal, with a delay. The picture, viewed from the z-transform domain is that of Figure 9.6.

We need to define $\downarrow M$ and $\uparrow M$.

Lemma 51 *In the time domain,* $\mathbf{y} = (\downarrow M)\mathbf{x}$ *has components* $\mathbf{y}(n) = \mathbf{x}(Mn)$. *In the frequency domain this is*

$$Y(\omega) = \frac{1}{M}\left[X(\frac{\omega}{M}) + X(\frac{\omega+2\pi}{M}) + \cdots + X(\frac{\omega+(M-1)2\pi}{M})\right].$$

The z-transform is

$$Y(z) = \frac{1}{M}\sum_{k=0}^{M-1} X(z^{1/M}e^{2\pi ik/M})$$

Lemma 52 *In the time domain,* $\mathbf{x} = (\uparrow M)\mathbf{y}$ *has components*

$$\mathbf{x}(k) = \begin{cases} \mathbf{y}(\frac{k}{M}), & \text{if } M \text{ divides } k \\ 0, & \text{otherwise} \end{cases}$$

In the frequency domain this is

$$X(\omega) = Y(M\omega).$$

The z-transform is

$$X(z) = Y(z^M).$$

The z-transform of $\mathbf{u} = (\downarrow M)(\uparrow M)\mathbf{x}$ *is*

$$U(z) = \frac{1}{M}\sum_{k=0}^{M-1} X(ze^{2\pi ik/M}).$$

Note that $(\uparrow M)(\downarrow M)$ is the identity operator, whereas $(\downarrow M)(\uparrow M)\mathbf{x}$ leaves every Mth element of $\mathbf{x}$ unchanged and replaces the rest by zeros.

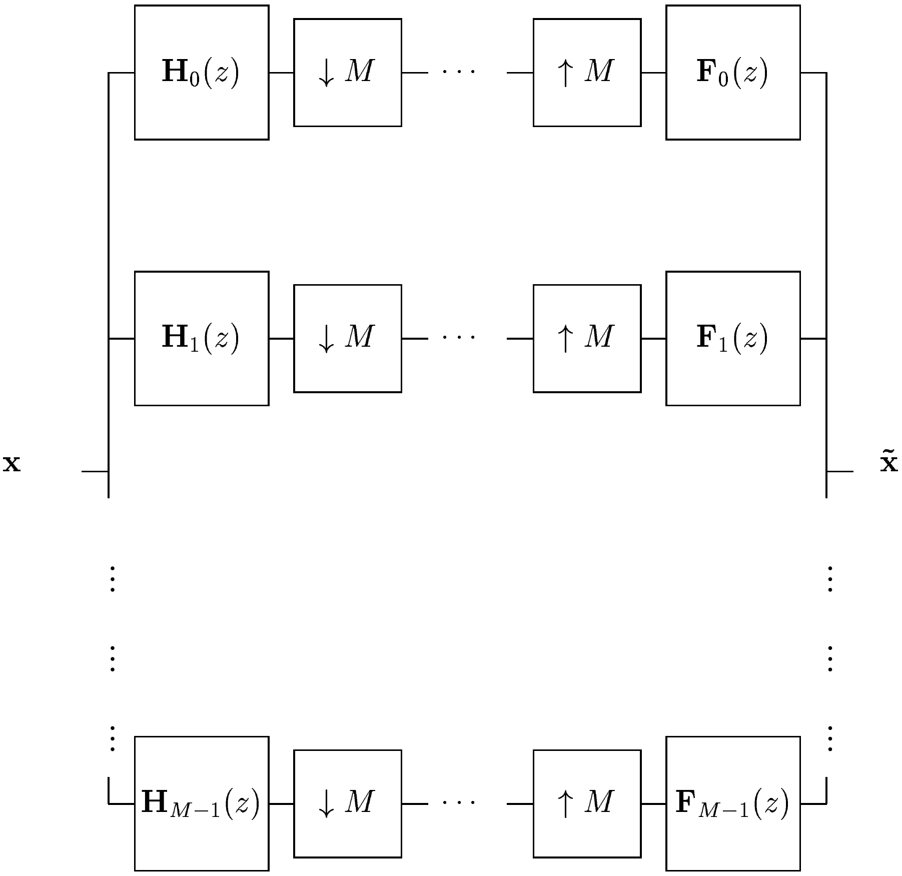

Figure 9.6: M-channel filter bank

The operator condition for perfect reconstruction with delay ℓ is

$$\sum_{j=0}^{M-1} \mathbf{F}_j(\uparrow 2)(\downarrow 2)\mathbf{H}_j = \mathbf{S}^\ell$$

where **S** is the shift. If we apply the operators on both sides of this requirement to a signal $\mathbf{x} = \{\mathbf{x}(n)\}$ and take the z-transform, we find

$$\frac{1}{M}\sum_{j=0}^{M-1} F_j(z) \sum_{k=0}^{M-1} H_j(zW^k)X(zW^k)$$

$$= z^{-\ell}X(z), \tag{9.87}$$

where $X(z)$ is the z-transform of $\mathbf{x}$, and $W = e^{-2\pi i/M}$. The coefficients of $X(zW^k)$ for $1 < k < M-1$ on the left-hand side of this equation are aliasing terms, due to the downsampling and upsampling. For perfect reconstruction of a general signal $X(z)$ these coefficients must vanish. Thus we have

Theorem 81 *An M channel filter bank gives perfect reconstruction when*

$$\text{No distortion}: \sum_{j=0}^{M-1} F_j(z)H_j(z) = Mz^{-\ell} \tag{9.88}$$

$$\text{Alias cancellation}: \sum_{j=0}^{M-1} F_j(z)H_j(zW^k) = 0, \qquad k = 1, \cdots, M-1 \tag{9.89}$$

In matrix form this reads

$$\begin{bmatrix} F_0(z) \\ F_1(z) \\ \vdots \\ F_{M-1}(z) \end{bmatrix} = \begin{bmatrix} H_0(z) & H_1(z) & \cdots & H_{M-1}(z) \\ H_0(zW) & H_1(zW) & \cdots & H_{M-1}(zW) \\ \vdots & \vdots & \vdots & \vdots \\ H_0(zW^{M-1}) & H_1(zW^{M-1}) & \cdots & H_{M-1}(zW^{M-1}) \end{bmatrix} = \begin{bmatrix} Mz^{-\ell} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where the $M \times M$ matrix is the *analysis modulation matrix* $\mathbf{H}_m(z)$. In the case $M = 2$ we could find a simple solution of the alias cancellation requirements (9.89) by defining the synthesis filters in terms of the analysis filters. However, this is not possible for general M and the design of these filter banks is more complicated. See Chapter 9 of Strang and Nguyen for more details.

Associated with M channel filter banks are M-band filters. The dilation and wavelet equations are

$$\phi_\ell(t) = M\sum_k \mathbf{h}_\ell(k)\phi(Mt-k), \qquad \ell = 0,1,\cdots,M-1.$$

Here $\mathbf{h}_\ell(k)$ is the finite impulse response vector of the FIR filter H_ℓ. Usually $\ell = 0$ is the dilation equation (for the scaling function $\phi_0(t)$ and $\ell = 1,\cdots,M-1$ are wavelet equations for $M-1$ wavelets $\phi_\ell(t)$, $\ell = 1,\cdots,M-1$. In the frequency domain the equations become

$$\hat{\phi}_\ell(\omega) = H(\frac{\omega}{M})\hat{\phi}_\ell(\frac{\omega}{M}), \qquad \ell = 0,1,\cdots,M-1,$$

and the iteration limit is

$$\hat{\phi}_\ell(\omega) = \prod_{k=1}^{\infty}\left(H(\frac{\omega}{M^k})\right)\hat{\phi}_\ell(0), \qquad \ell = 0,1,\cdots,M-1,$$

assuming that the limit exists and $\hat{\phi}_\ell(0)$ is well defined. For more information about these M-band wavelets and the associated multiresolution structure, see the book by Burrus, Gopinath and Rao.

9.5.2 Multifilters and Multiwavelets

Next we go back to filters corresponding to the case $M = 2$, but now we let the input vector $\mathbf{x}$ be an $r \times \infty$ input matrix. Thus each component $\mathbf{x}(n)$ is an r-vector $\mathbf{x}_j(n), \quad j = 1,\cdots,r$. Instead of analysis filters H_0, H_1 we have analysis multifilters $\mathbf{H}_0, \mathbf{H}_1$, each of which is an $r \times r$ matrix of filters, e.g., $\mathbf{H}_0^{(j,k)}, \quad j,k = 1,\cdots r$. Similarly we have synthesis multifilters $\mathbf{F}_0, \mathbf{F}_1$, each of which is an $r \times r$ matrix of filters.

Formally, part of the theory looks very similar to the scalar case. Thus the z-transform of a multifilter $\mathbf{H}$ is $\mathbf{H}(z) = \sum_k \mathbf{h}(k)z^{-k}$ where $\mathbf{h}(k)$ is the $r \times r$ matrix of filter coefficients. In fact we have the following theorem (with the same proof).

Theorem 82 *A multifilter gives perfect reconstruction when*

$$\text{No distortion}: \mathbf{F}_0(z)\mathbf{H}_0(z) + \mathbf{F}_1(z)\mathbf{H}_1(z) = 2z^{-\ell}\mathbf{I} \qquad (9.90)$$

$$\text{Alias cancellation}: \mathbf{F}_0(z)\mathbf{H}_0(-z) + \mathbf{F}_1(z)\mathbf{H}_1(-z) = \mathbf{0}. \qquad (9.91)$$

Here, all of the matrices are $r \times r$. We can no longer give a simple solution to the alias cancellation equation, because $r \times r$ matrices do not, in general, commute.

There is a corresponding theory of multiwavelets. The dilation equation is

$$\Phi(t) = 2\sum_k \mathbf{h}_0(k)\Phi(2t-k),$$

where $\Phi_j(t), \quad j = 1, \cdots, r$ is a vector of r scaling functions. the wavelet equation is

$$\mathbf{w}(t) = 2\sum_k \mathbf{h}_1(k)\mathbf{w}(2t-k),$$

where $\mathbf{w}_j(t), \quad j = 1, \cdots, r$ is a vector of r wavelets.

A simple example of a multiresolution analysis is “Haar’s hat”. Here the space V_0 consists of piecewise linear (discontinuous) functions. Each such function is linear between integer gridpoints $t_j = j$ and right continuous at the gridpoints: $f(j) = f(j+0)$. However, in general $f(j) \neq f(j-0)$. Each such function is uniquely determined by the 2-component input vector $(f(j), f(j+1-0)) = (a_j, b_j)$, $j = 0, \pm 1, \cdots$. Indeed, $f(t) = (b_j - a_j)(t-j) + a_j$ for $j \le t < j+1$. We can write this representation as

$$f(t) = \frac{b_j + a_j}{2}\phi_1(t-j) + \frac{b_j - a_j}{2}\phi_2(t-j), \qquad j \le t < j+1,$$

where $\frac{b_j+a_j}{2}$ is the *average* of f in the interval $j \le t < j+1$ and $b_j - a_j$ is the *slope*. Here,

$$\phi_1(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases} \qquad \phi_2(t) = \begin{cases} 2t-1 & 0 \le t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\phi_1(t)$ is just the box function. Note further that the integer translates of the two scaling functions $\phi_1(t+k)$, $\phi_2(t+\ell)$ are mutually orthogonal and form an ON basis for V_0. The same construction goes over if we halve the interval. The dilation equation for the box function is (as usual)

$$\phi_1(t) = \phi_1(2t) + \phi_1(2t-1).$$

You can verify that the dilation equation for $\phi_2(t)$ is

$$\phi_2(t) = \frac{1}{2}\left[\phi_2(2t) + \phi_2(2t-1) - \phi_1(2t) + \phi_1(2t-1)\right].$$

They go together in the *matrix dilation equation*

$$\begin{bmatrix} \phi_1(t) \\ \phi_2(t) \end{bmatrix} = \begin{bmatrix} 1, & 0 \\ -\frac{1}{2}, & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \phi_1(2t) \\ \phi_2(2t) \end{bmatrix} + \begin{bmatrix} 1, & 0 \\ \frac{1}{2}, & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \phi_1(2t-1) \\ \phi_2(2t-1) \end{bmatrix}. \tag{9.92}$$

See Chapter 9 of Strang and Nguyan, and the book by Burrus, Gopinath and Guo for more details.

9.6 Finite Length Signals

In our previous study of discrete signals in the time domain we have usually assumed that these signals were of infinite length, an we have designed filters and passed to the treatment of wavelets with this in mind. This is a good assumption for files of indefinite length such as audio files. However, some files (in particular video files) have a fixed length L. How do we modify the theory, and the design of filter banks, to process signals of fixed finite length L? We give a very brief discussion of this. Many more details are found in the text of Strang and Nguyen.

Here is the basic problem. The input to our filter bank is the finite signal $\mathbf{x}(0), \cdots \mathbf{x}(L-1)$, and nothing more. What do we do when the filters call for values $\mathbf{X}(n)$ where n lies outside that range? There are two basic approaches. One is to redesign the filters (so called boundary filters) to process only blocks of length L. We shall not treat this approach in these notes. The other is to embed the signal of length L as part of an infinite signal (in which no additional information is transmitted) and to process the extended signal in the usual way. Here are some of the possibilities:

1. *Zero-padding*, or *constant-padding*. Set $\mathbf{x}(n) = c$ for all $n < 0$ or $n \geq L$. Here c is a constant, usually 0. If the signal is a sampling of a continuous function, then zero-padding ordinarily introduces a discontinuity.

2. Extension by periodicity (*wraparound*). We require $\mathbf{x}(n) = \mathbf{x}(m)$ if $n = m, \bmod L$, i.e., $\mathbf{x}(n) = \mathbf{x}(k)$ for $k = 0, 1, \cdots, L-1$ if $n = aL+k$ for some integer a. Again this ordinarily introduces a discontinuity. However, Strang and Nguyen produce some images to show that wraparound is ordinarily superior to zero-padding for image quality, particularly if the image data is nearly periodic at the boundary.

3. Extension by reflection. There are two principle ways this is done. The first is called **whole-point symmetry**, or $\mathbf{W}$. We are given the finite signal

$\mathbf{x}(0), \cdots \mathbf{x}(L-1)$. To extend it we reflect at position 0. Thus, we define $\mathbf{x}(-1) = \mathbf{x}(1), \mathbf{x}(-2) = \mathbf{x}(2), \cdots, \mathbf{x}(-[L-2]) = \mathbf{x}(L-2)$. This defines $\mathbf{x}(n)$ in the $2L-2$ strip $n = -L+2, \cdots, L-1$. Note that the values of $\mathbf{x}(0)$ and $\mathbf{x}(L-1)$ each occur once in this strip, whereas the values of $\mathbf{x}(1), \cdots, \mathbf{x}(L-2)$ each occur twice. Now $\mathbf{x}(n)$ is defined for general n by $2L-2$ periodicity. Thus whole-point symmetry is a special case of wraparound, but the periodicity is $2L-2$, not L. This is sometimes referred to as a (1,1) extension, since neither endpoint is repeated.

- The second symmetric extension method is called **half-point symmetry**, or $\mathbf{H}$. We are given the finite signal $\mathbf{x}(0), \cdots \mathbf{x}(L-1)$. To extend it we reflect at position $-\frac{1}{2}$, halfway between 0 and -1. Thus, we define $\mathbf{x}(-1) = \mathbf{x}(0), \mathbf{x}(-2) = \mathbf{x}(1), \cdots, \mathbf{x}(-L) = \mathbf{x}(L-1)$. This defines $\mathbf{x}(n)$ in the $2L$ strip $n = -L, \cdots, L-1$. Note that the values of $\mathbf{x}(0)$ and $\mathbf{x}(L-1)$ each occur twice in this strip, as do the values of $\mathbf{x}(1), \cdots, \mathbf{x}(L-2)$. Now $\mathbf{x}(n)$ is defined for general n by $2L$ periodicity. Thus **H** is again a special case of wraparound, but the periodicity is $2L$, not L, or $2L-2$. This is sometimes referred to as a (2,2) extension, since both endpoints are repeated.

Strang and Nguyen produce some images to show that symmetric extension is modestly superior to wraparound for image quality. If the data is a sampling of a differentiable function, symmetric extension maintains continuity at the boundary, but introduces a discontinuity in the first derivative.

9.6.1 Circulant Matrices

All of the methods for treating finite length signals of length L introduced in the previous section involve extending the signal as infinite and periodic. For wraparound, the period is L; for whole-point symmetry $\mathbf{W}$ the period is $2L-2$; for half-point symmetry $\mathbf{H}$ it is $2L$. To take advantage of this structure we modify the definitions of the filters so that they exhibit this same periodicity. We will adopt the notation of Strang and Nguyen and call this period L, (with the understanding that this number is the period of the underlying data: $\tilde{L}$, $2\tilde{L}-2$ or $2\tilde{L}$). Then the data can be considered as as a repeating L-tuple and the filters map repeating L-tuples to repeating L-tuples. Thus for passage from the time domain to the frequency domain, we are, in effect, using the discrete Fourier transform (DFT), base L.

For infinite signals the matrices of FIR filters $\mathbf{H}$ are Toeplitz, the filter action is given by convolution, and this action is diagonalized in the frequency domain as multiplication by the Fourier transform of the finite impulse response vector of $\mathbf{H}$. There are perfect $L \times L$ analogies for Toeplitz matrices. These are the *circulant matrices*. Thus, the infinite signal in the time domain becomes an L-periodic signal, the filter action by Toeplitz matrices becomes action by $L \times L$ circulant matrices and the finite Fourier transform to the frequency domain becomes the DFT, base L. Implicitly, we have worked out most of the mathematics of this action in Chapter 5. We recall some of this material to link with the notation of Strang and Nguyen and the concepts of filter bank theory.

Recall that the infinite matrix FIR filter $\mathbf{H}$ can be expressed in the form $\mathbf{H} = \sum_{k=0}^{N} \mathbf{h}(k)\mathbf{S}^k$ where $\mathbf{h}(n)$ is the impulse response vector and $\mathbf{S}$ is the infinite shift matrix $\mathbf{S}\mathbf{x}(n) = \mathbf{x}(n-1)$. If $N < L$ we can define the action of $\mathbf{H}$ (on data consisting of repeating L-tuples) by restriction. Thus the shift matrix becomes the $L \times L$ *cyclic permutation matrix* $\mathbf{S}_L$ defined by $\mathbf{S}_L\mathbf{x}(n) = \mathbf{x}(n-1), \quad \mod L$. For example:

$$\mathbf{S}_4\mathbf{x} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}(0) \\ \mathbf{x}(1) \\ \mathbf{x}(2) \\ \mathbf{x}(3) \end{bmatrix} = \begin{bmatrix} \mathbf{x}(3) \\ \mathbf{x}(0) \\ \mathbf{x}(1) \\ \mathbf{x}(2) \end{bmatrix}.$$

The matrix action of $\mathbf{H}$ on repeating L-tuples becomes

$$\mathbf{H}_L = \sum_{n=0}^{N} \mathbf{h}(n)(\mathbf{S}_L)^n.$$

This is an instance of a circulant matrix.

Definition 36 *An $L \times L$ matrix $\mathbf{A}$ is called a circulant if all of its diagonals (main, sub and super) are constant and the indices are interpreted mod L. Thus, there is an L-vector vector $a_(k)$ such that $A_{\ell,k} = a(\ell - k)$ mod L.*

Example 12

$$\begin{pmatrix} 1 & 5 & 3 & 2 \\ 2 & 1 & 5 & 3 \\ 3 & 2 & 1 & 5 \\ 5 & 3 & 2 & 1 \end{pmatrix}.$$

Recall that the column vector

$$X = (X[0], X[1], X[2], \cdots, X[L-1])$$

is the *Discrete Fourier transform* (DFT) of $\mathbf{x} = \{\mathbf{x}[n],\ n = 0, 1, \cdots, L-1\}$ if it is given by the matrix equation $X = \mathcal{F}_L \mathbf{x}$ or

$$\begin{pmatrix} X[0] \\ X[1] \\ X[2] \\ \vdots \\ X[L-1] \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{L-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(L-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{L-1} & \omega^{2(L-1)} & \cdots & \omega^{(L-1)(L-1)} \end{pmatrix} \begin{pmatrix} \mathbf{x}[0] \\ \mathbf{x}[1] \\ \mathbf{x}[2] \\ \vdots \\ \mathbf{x}[L-1] \end{pmatrix} \tag{9.93}$$

where $\omega = \overline{W} = e^{-2\pi i/L}$. Thus,

$$X[k] = \tilde{\mathbf{x}}(k) = \sum_{n=0}^{L-1} \mathbf{x}[n]\overline{W}^{nk} = \sum_{n=0}^{L-1} \mathbf{x}[n] e^{-2\pi i n k/L}.$$

Here $\mathcal{F}_L = \overline{\mathbf{F}_L}$ is an $L \times L$ matrix. The inverse relation is the matrix equation $\mathbf{x} = \mathcal{F}_L^{-1} X$ or

$$\begin{pmatrix} \mathbf{x}[0] \\ \mathbf{x}[1] \\ \mathbf{x}[2] \\ \vdots \\ \mathbf{x}[L-1] \end{pmatrix} = \frac{1}{L} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \bar{\omega} & \bar{\omega}^2 & \cdots & \bar{\omega}^{L-1} \\ 1 & \bar{\omega}^2 & \bar{\omega}^4 & \cdots & \bar{\omega}^{2(L-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{\omega}^{L-1} & \bar{\omega}^{2(L-1)} & \cdots & \bar{\omega}^{(L-1)(L-1)} \end{pmatrix} \begin{pmatrix} X[0] \\ X[1] \\ X[2] \\ \vdots \\ X[L-1] \end{pmatrix}, \tag{9.94}$$

or

$$\mathbf{x}[n] = \frac{1}{L} \sum_{k=0}^{L-1} X[k] W^{nk} = \frac{1}{L} \sum_{k=0}^{L-1} X[k] e^{2\pi i n k/L},$$

where $\mathcal{F}^{-1} = \frac{1}{L}\mathbf{F}_L$ and $\overline{W} = \omega = e^{-2\pi i/L}, W = \bar{\omega} = \omega^{-1} = e^{2\pi i/L}$.

Note that

$$\overline{\mathbf{F}_L}\mathbf{F}_L = L\mathbf{I}, \qquad \overline{\mathbf{F}_L}^{-1} = \frac{1}{L}\mathbf{F}_L.$$

For $\mathbf{x}, \mathbf{y} \in P_L$ (the space of repeating L-tuples) we define the *convolution* $\mathbf{x} * \mathbf{y} \in P_L$ by

$$\mathbf{x} * \mathbf{y}[n] = \mathbf{z}[n] \equiv \sum_{m=0}^{L-1} \mathbf{x}[m]\mathbf{y}[n-m].$$

Then $Z[k] = X[k]Y[k]$.

Now note that the DFT of $\mathbf{S}_l^n\mathbf{x}[m]$ is $\hat{\mathbf{S}}_L^n\mathbf{x}[k]$ where

$$\hat{\mathbf{S}}_L^n\mathbf{x}[k] = \sum_{m=0}^{L-1} \mathbf{S}_L^n\mathbf{x}[m]e^{-2\pi imk/L} = \sum_{m=0}^{L-1} \mathbf{x}[m-n]e^{-2\pi imk/L}$$

$$= \overline{W}^{nk}\sum_{m=0}^{L-1} \mathbf{x}[m]e^{-2\pi imk/L} = \overline{W}^{nk}X[k] = \overline{W}^{nk}\hat{\mathbf{x}}[k].$$

Thus,

$$\hat{\mathbf{H}}_L\mathbf{x}[k] = \left(\sum_{n=0}^{N} \mathbf{h}(n)\hat{\mathbf{S}}_L^n\mathbf{x}\right)[k] = \left(\sum_{n=0}^{N} \mathbf{h}(n)\overline{W}^{nk}\right)\hat{\mathbf{x}}[k]$$

$$= \hat{\mathbf{h}}[k]\hat{\mathbf{x}}[k].$$

In matrix notation, this reads

$$\hat{\mathbf{H}}_L\mathbf{x} = \overline{\mathbf{F}_L}\mathbf{H}_L\mathbf{x} = \mathbf{D}_H\overline{\mathbf{F}_L}\mathbf{x}$$

where $\mathbf{D}_H$ is the $L \times L$ diagonal matrix

$$(\mathbf{D}_H)_{jk} = \hat{\mathbf{h}}[k]\delta_{jk}.$$

Since $\mathbf{x}$ is arbitrary we have $\overline{\mathbf{F}_L}\mathbf{H}_L = \mathbf{D}_H\overline{\mathbf{F}_L}$ or

$$\mathbf{D}_H = \mathbf{F}_L^{-1}\mathbf{H}_L\mathbf{F}_L.$$

This is the exact analog of the convolution theorem for Toeplitz matrices in frequency space. It says that circulant matrices are diagonal in the DFT frequency space, with diagonal elements that are the DFT of the impulse response vector $\mathbf{h}$.

9.6.2 Symmetric Extension for Symmetric Filters

As Strang and Nguyen show, symmetric extension of signals is usually superior to wraparound alone, because it avoids introducing jumps in samplings of continuous signals.

However, symmetric extension, either **W** or **H**, introduces some new issues. Our original input is L numbers $\mathbf{x}(0), \cdots, \mathbf{x}(L-1)$. We extend the signal $\mathbf{x}$, by **W** to get a $2(L-1)$-periodic signal, or by **H** to get a $2L$-periodic signal. Then we filter the extended signal **Ex** by a low pass filter H_0 and, separately, by a high pass filter H_1, each with filter length $N < L$. Following this we downsample

$\downarrow 2$ the outputs from the analysis filters. The outputs from each downsampled analysis filter will contain either $L-1$ elements (**W**) or L elements (**H**). From this collection of $2(L-1)$ or $2L$ downsampled elements we must be able to find a restricted subset of L independent elements, from which we can reconstruct the original input of L numbers via the synthesis filters.

An important strategy to make this work is to assure that the downsampled signals $(\downarrow 2)\mathbf{HEx}$ are symmetric so that about half of the elements are obviously redundant. Then the selection of L independent elements (about half from the low pass downsampled output and about half from the high pass downsampled output) becomes much easier. The following is a successful strategy. We choose the filters H to be symmetric, i.e., $\mathbf{h}(N-n)=\mathbf{h}(n)$.

Definition 37 *If $\mathbf{H}$ is a symmetric FIR filter and N is even, so that the impulse response vector $\mathbf{h}$ has odd length, then $\mathbf{H}$ is called a $\mathbf{W}$ filter. It is symmetric about its midpoint $N/2$ and $\mathbf{h}(N/2)$ occurs only once. If N is odd, then $\mathbf{H}$ is called an $\mathbf{H}$ filter. It is symmetric about $N/2$, a gap midway between two repeated coefficients.*

Lemma 53 *If $\mathbf{H}$ is a $\mathbf{W}$ filter and $\mathbf{Ex}$ is a $\mathbf{W}$ extension of $\mathbf{x}$, then $\mathbf{y}=\mathbf{HEx}$ is a $\mathbf{W}$ extension and $(\downarrow 2)\mathbf{y}$ is symmetric. Similarly, if $\mathbf{H}$ is an $\mathbf{H}$ filter and $\mathbf{Ex}$ is a $\mathbf{H}$ extension of $\mathbf{x}$, then $\mathbf{y}=\mathbf{HEx}$ is a $\mathbf{W}$ extension and $(\downarrow 2)\mathbf{y}$ is symmetric.*

PROOF: Suppose $\mathbf{H}$ is a $\mathbf{W}$ filter and $\mathbf{Ex}$ is a $\mathbf{W}$ extension of $\mathbf{x}$. Thus $\mathbf{h}(N-n)=\mathbf{h}(n)$ where N is even, and $\mathbf{x}(m)=\mathbf{x}(-m)$ with $\mathbf{x}(m+\tilde{L})=\mathbf{x}(m)$ for $\tilde{L}=2(L-1)$. Now set $\mathbf{y}(n)=\sum_m \mathbf{h}(n-m)\mathbf{x}(m)$. We have

$$\mathbf{y}(N-m)=\sum_m \mathbf{h}(N-n-m)\mathbf{x}(m)=\sum_m \mathbf{h}(n+m)\mathbf{x}(m)$$

$$=\sum_m \mathbf{h}(n-m)\mathbf{x}(-m)=\sum_m \mathbf{h}(n-m)\mathbf{x}(m)=\mathbf{y}(n).$$

Also,

$$\mathbf{y}(N+\tilde{L})=\sum_m \mathbf{h}(n+\tilde{L}-m)\mathbf{x}(m)=\sum_m \mathbf{h}(n-[m-\tilde{L}])\mathbf{x}(m)$$

$$=\sum_k \mathbf{h}(n-k)\mathbf{x}(k+\tilde{L})=\sum_k \mathbf{h}(n-k)\mathbf{x}(k)=\mathbf{y}(n).$$

Then $(\downarrow 2)\mathbf{y}(k)=\mathbf{y}(2k)$ so

$$(\downarrow 2)\mathbf{y}(\frac{N}{2}-k)=\mathbf{y}(N-2k)=\mathbf{y}(2k)=(\downarrow 2)\mathbf{y}(k).$$

Similarly, suppose **H** is a **H** filter and **Ex** is a **H** extension of $\mathbf{x}$. Thus $\mathbf{h}(N-n)=\mathbf{h}(n)$ where N is odd, and $\mathbf{x}(m)=\mathbf{x}(-m-1)$ with $\mathbf{x}(m+\tilde{L})=\mathbf{x}(m)$ for $\tilde{L}=2L$. Now set $\mathbf{y}(n)=\sum_m \mathbf{h}(n-m)\mathbf{x}(m)$. We have

$$\mathbf{y}(N-m)=\sum_m \mathbf{h}(N-n-m)\mathbf{x}(m)=\sum_m \mathbf{h}(n+m)\mathbf{x}(m)$$

$$=\sum_m \mathbf{h}(n-m)\mathbf{x}(-m)=\sum_m \mathbf{h}(n-m)\mathbf{x}(m-1)=\mathbf{y}(n-1).$$

Also,

$$\mathbf{y}(N+\tilde{L})=\sum_m \mathbf{h}(n+\tilde{L}-m)\mathbf{x}(m)=\sum_m \mathbf{h}(n-[m-\tilde{L}])\mathbf{x}(m)$$

$$=\sum_k \mathbf{h}(n-k)\mathbf{x}(k+\tilde{L})=\sum_k \mathbf{h}(n-k)\mathbf{x}(k)=\mathbf{y}(n).$$

Then $(\downarrow 2)\mathbf{y}(k)=\mathbf{y}(2k)$ so

$$(\downarrow 2)\mathbf{y}(\frac{N-1}{2}-k)=\mathbf{y}(N-1-2k)=\mathbf{y}(2k)=(\downarrow 2)\mathbf{y}(k).$$

Q.E.D.

REMARKS:

1. Though we have given proofs only for the symmetric case, the filters can also be antisymmetric, i.e., $\mathbf{h}(N-n)=-\mathbf{h}(n)$. The antisymmetry is inherited by $\mathbf{y}(k)$ and $(\downarrow 2)\mathbf{y}(k)$.

2. For N odd (**W** filter) if the low pass FIR filter is real and symmetric and the high pass filter is obtained via the alternating flip, then the high pass filter is also symmetric. However, if N is even (**H** filter) and the low pass FIR filter is real and symmetric and the high pass filter is obtained via the alternating flip, then the high pass filter is antisymmetric.

3. The pairing of **W** filters with **W** extensions of signals (and the pairing of **H** filters with **H** extensions of signals) is important for the preceding result. Mixing the symmetry types will result in downsampled signals without the desired symmetry.

4. The exact choice of the elements of the restricted set, needed to reconstitute the original L-element signal depend on such matters as whether L is odd or even. Thus, for a **W** filter (N even) and a **W** extension $\mathbf{E}$ with L even, then since $L-1$ is odd $(\downarrow 2)\mathbf{y}$ must be a $(1,2)$ extension (one endpoint occurs once and one is repeated) so we can choose $1+\frac{L-2}{2}=\frac{L}{2}$ independent elements from each of the upper and lower channels. However, if L is odd then $L-1$ is even and $(\downarrow 2)\mathbf{y}$ must be a $(1,1)$ extension. In this case the independent components are $2+\frac{L-3}{2}$. Thus by correct centering we can choose $\frac{L+1}{2}$ elements from one channel and $\frac{L-1}{2}$ from the other.

5. For a symmetric low pass **H** filter (N odd) and an **H** extension $\mathbf{E}$ with L even, then $(\downarrow 2)\mathbf{y}$ must be a $(1,1)$ extension so we can choose $2+\frac{L-2}{2}=\frac{L}{2}+1$ independent elements from the lower channel. The antisymmetry in the upper channel forces the two endpoints to be zero, so we can choose $\frac{L-2}{2}=\frac{L}{2}-1$ independent elements. However, if L is odd then $(\downarrow 2)\mathbf{y}$ must be a $(1,2)$ extension. In this case the independent components in the lower channel are $2+\frac{L-3}{2}$. The antisymmetry in the upper channel forces one endpoint to be zero. Thus by correct centering we can choose $\frac{L-1}{2}$ independent elements from the upper channel.

6. Another topic related to the material presented here is the Discrete Cosine Transform (DCT). Recall that the Discrete Fourier Transform (DFT) essentially maps the data $f(0), f(1), \cdots, f(L-1)$ on the interval $[0, L-1]$ to equally spaced points around the unit circle. On the circle the points $L-1$ and 0 are adjoining. Thus the DFT of samples of a continuous function f on an interval can have an artificial discontinuity when passing from $L-1$ to 0 on the circle. This leads to the Gibb's phenomenon and slow convergence. One way to fix this is to use the basic idea behind the Fourier cosine transform and to make a symmetric extension g of f to the interval of length $2L$:
$$g(n) = \begin{cases} f(n) & n = 0, \cdots, L-1 \\ f(-n-1) & n = -L, -L+1, \cdots, -1 \end{cases}$$
Now $g(-L) = g(L-1)$, so that if we compute the DFT of g on an interval of length $2L$ we will avoid the discontinuity problems and improve convergence. Then at the end we can restrict to the interval $[0, L-1]$.

7. More details can be found in Chapter 8 of Strang and Nguyen.

Chapter 10

Some Applications of Wavelets

10.1 Image compression

A typical image consists of a rectangular array of 256×256 pixels, each pixel coded by 24 bits. In contrast to an audio signal, this signal has a fixed length. The pixels are transmitted one at a time, starting in the upper left-hand corner of the image and ending with the lower right. However for image processing purposes it is more convenient to take advantage of the $2D$ geometry of the situation and consider the image not as a linear time sequence of pixel values but as a geometrical array in which each pixel is assigned its proper location in the image. Thus the finite signal is 2-dimensional: $\mathbf{x}(n_1, n_2)$ where $0 \leq n_i < 2^8$. We give a very brief introduction to subband coding for the processing of these images. Much more detail can be found in chapters 9-11 of Strang and Nguyen.

Since we are in 2 dimensions we need a $2D$ filter $\mathbf{H}$

$$\mathbf{Hx}(n_1, n_2) = \mathbf{y}(n_1, n_2) = \sum_{k_1, k_2} \mathbf{h}(k_1, k_2)\mathbf{x}(n_1 - k_1, n_2 - k_2).$$

This is the $2D$ convolution $\mathbf{y} = \mathbf{Hx}$. In the frequency domain this reads

$$Y(\omega_1, \omega_2) = H(\omega_1, \omega_2)X(\omega_1, \omega_2) = \left(\sum_{k_1, k_2} \mathbf{h}(k_1, k_2)e^{-i(k_1\omega_1 + k_2\omega_2)}\right) \times$$

$$\left(\sum_{n_1, n_2} \mathbf{x}(n_1, n_2)e^{-i(n_1\omega_1 + n_2\omega_2)}\right),$$

with a similar expression for the z-transform. The frequency response is 2π-periodic in each of the variables ω_j and the frequency domain is the square $-\pi \leq$

$\omega_j < \pi$. We could develop a truly $2D$ filter bank to process this image. Instead we will take the easy way out and use *separable* filters, i.e., products of $1D$ filters. We want to decompose the image into low frequency and high frequency components in each variable n_1, n_2 separately, so we will use four separable filters, each constructed from one of our $1D$ pairs $\mathbf{h}_{low}, \mathbf{h}_{high}$ associated with a wavelet family:

$$\mathbf{h}_0(n_1,n_2) = \mathbf{h}_{low}(n_1)\mathbf{h}_{low}(n_2), \qquad \mathbf{h}_2(n_1,n_2) = \mathbf{h}_{high}(n_1)\mathbf{h}_{low}(n_2),$$

$$\mathbf{h}_1(n_1,n_2) = \mathbf{h}_{low}(n_1)\mathbf{h}_{high}(n_2), \qquad \mathbf{h}_3(n_1,n_2) = \mathbf{h}_{high}(n_1)\mathbf{h}_{high}(n_2).$$

The frequency responses of these filters also factor. We have

$$H_2(\omega_1,\omega_2) = H_{high}(\omega_1)H_{low}(\omega_2),$$

etc. After obtaining the outputs $\mathbf{y}_j(n_1,n_2)$ from each of the four filters $\mathbf{H}_j$ we downsample to get $(\downarrow [2,2])\mathbf{y}(n_1,n_2) = \mathbf{y}_j(2n_1,2n_2)$. Thus we keep one sample out of four for each analysis filter. This means that we have exactly as many pixels as we started with (256×256), but now they are grouped into four (128×128) arrays. Thus the analyzed image is the same size as the original image but broken into four equally sized squares: LL (upper left), HL (upper right), LH (lower left), and HH (lower right). Here HL denotes the filter that is high pass on the n_1 index and low pass on the n_2 index, etc.

A straight-forward z-transform analysis shows that this is a perfect reconstruction $2D$ filter bank provided the factors $\mathbf{h}_{low}, \mathbf{h}_{high}$ define a $1D$ perfect reconstruction filter bank. the synthesis filters can be composed from the analogous synthesis filters for the factors. Upsampling is done in both indices simultaneously: $(\uparrow [2,2])\mathbf{y}(2n_1,2n_2) = \mathbf{y}(n_1,n_2)$ for the even-even indices. $(\uparrow [2,2])\mathbf{y}(m_1,m_2) = 0$ for m_1, m_2 even-odd, odd-even or odd-odd.

At this point the analysis filter bank has decomposed the image into four parts. LL is the analog of the low pass image. HL, LH and HH each contain high frequency (or difference) information and are analogs of the wavelet components. In analogy with the $1D$ wavelet transform, we can now leave the (128×128) wavelet subimages HL, LH and HH unchanged, and apply our $2D$ filter bank to the (128×128) LL subimage. Then this block in the upper left-hand corner of the analysis image will be replaced by four 64×64 blocks L'L', H'L', L'H' and H'H', in the usual order. We could stop here, or we could apply the filter bank to L'L' and divide it into four 32×32 pixel blocks L"L", H"L", L"H" and H"H". Each iteration adds a net three additional subbands to the analyzed image. Thus one pass through the filter bank gives 4 subbands, two passes give 7, three passes yield 10 and four yield 13. Four or five levels are common. For a typical analyzed image, most of the signal energy is in the low pass image in the small square in the upper left-hand corner. It appears as a bright but blurry miniature facsimile of the original image. The various wavelet subbands have little energy and are relatively dark.

If we run the analyzed image through the synthesis filter bank, iterating an appropriate number of times, we will reconstruct the original signal. However, the usual reason for going through this procedure is to process the image before reconstruction. The storage of images consumes a huge number of bits in storage devices; compression of the number of bits defining the image, say by a factor of 50, has a great impact on the amount of storage needed. Transmission of images over data networks is greatly speeded by image compression. The human visual system is very relevant here. One wants to compress the image in ways that are not apparent to the human eye. The notion of "barely perceptible difference" is important in multimedia, both for vision and sound. In the original image each pixel is assigned a certain number of bits, 24 in our example, and these bits determine the color and intensity of each pixel in discrete units. If we increase the size of the units in a given subband then fewer bits will be needed per pixel in that subband and fewer bits will need to be stored. This will result in a loss of detail but may not be apparent to the eye, particularly in subbands with low energy. This is called quantization. The compression level, say 20 to 1, is mandated in advance. Then a bit allocation algorithm decides how many bits to allocate to each subband to achieve that over-all compression while giving relatively more bits to high energy parts of the image, minimizing distortion, etc. (This is a complicated subject.) Then the newly quantized system is *entropy coded*. After quantization there may be long sequences of bits that are identical, say 0. Entropy coding replaces that long strong of 0s by the information that all of the bits from location a to location b are 0. The point is that this information can be coded in many fewer bits than were contained in the original sequence of 0s. Then the quantized and coded file is stored or transmitted. Later the compressed file is processed by the synthesizing filter bank to produce an image.

There are many other uses of wavelet based image processing, such as edge detection. For edge detection one is looking for regions of rapid change in the image and the wavelet subbands are excellent for this. Noise will also appear in the wavelet subbands and a noisy signal could lead to false positives by edge detection algorithms. To distinguish edges from noise one can use the criteria that an edge should show up at all wavelet levels. Your text contains much more information about all these matters.

10.2 Thresholding and Denoising

Suppose that we have analyzed a signal down several levels using the DWT. If the wavelets used are appropriate for the signal, most of the energy of the signal ($\sum_k |a_j,k|^2 = \sum_k (|a_{j-1,k}|^2 + |b_{j-1,k}|^2)$ at level j will be associated with just a few coefficients. The other coefficients will be small in absolute value. The basic idea behind thresholding is to zero out the small coefficients and to yield an economical representation of the signal by a few large coefficients. At each wavelet level one chooses a threshold $\delta > 0$. Suppose $x_j(t)$ is the projection of the signal at that level. There are two commonly used methods for thresholding. For *hard thresholding* we modify the signal according to

$$y_{\text{hard},j}(t) = \begin{cases} x_j(t), & \text{for } |x_j(t)| > \delta \\ 0, & \text{for } |x_j(t)| \leq \delta. \end{cases}$$

Then we synthesize the modified signal. This method is very simple but does introduce discontinuities. For *soft thresholding* we modify the signal continuously according to

$$y_{\text{soft},j}(t) = \begin{cases} \text{sign}(x(t))(x_j(t) - \delta), & \text{for } |x_j(t)| > \delta \\ 0, & \text{for } |x_j(t)| \leq \delta. \end{cases}$$

Then we synthesize the modified signal. This method doesn't introduce discontinuities. Note that both methods reduce the overall signal energy. It is necessary to know something about the characteristics of the desired signal in advance to make sure that thresholding doesn't distort the signal excessively.

Denoising is a procedure to recover a signal that has been corrupted by noise. (Mathematically, this could be Gaussian white noise $N(0,1)$) It is assumed that the basic characteristics of the signal are known in advance and that the noise power is much smaller than the signal power. A familiar example is static in a radio signal. We have already looked at another example, the noisy Doppler signal in Figure 7.5.

The idea is that when the signal plus noise is analyzed via the DWT the essence of the basic signal shows up in the low pass channels. Most of the noise is captured in the differencing (wavelet) channels. You can see that clearly in Figures 7.5, 7.6, 7.7. In a channel where noise is evident we can remove much of it by soft thresholding. Then the reconstituted output will contain the signal with less noise corruption.