Chapter 2

Filters

This chapter comes between the Haar example of Chapter 1 and the full development of filter banks (leading to wavelets). We decided to collect other definitions that belong to this circle of ideas. Here is an indication of our plan:

Basic filters: Ideal filters and then FIR filter design
Basic tools: Fourier methods and functional analysis
Bases and frames: A matrix $\boldsymbol{T}$ has a two-sided or only a one-sided inverse
Integral transforms: windows in time-frequency, wavelets in time-scale.

Signal processing is an enormous subject. The input signal can arrive in many forms: continuous time, discrete time, finite time. It can be processed in many ways. Our greatest interest is a signal $x(n)$ in discrete time that is processed by a linear time-invariant operator. *If the input is shifted in time then the output is equally shifted.* These operators are ***filters*** — the fundamental actors in signal processing.

Filters can be expressed in three domains: n, ω, z. In each domain the filter is a multiplication:

(n) : Multiplication by a ***Toeplitz matrix*** with $\boldsymbol{h}(n)$ on the nth diagonal.
(ω) : Multiplication by the ***frequency response*** $H(e^{j\omega}) = \sum \boldsymbol{h}(n)e^{-j\omega n}$.
(z) : Multiplication by the ***transfer function*** $H(z) = \sum \boldsymbol{h}(n)z^{-n}$.

We want to explain these three forms and the connections between them. *If we emphasize the matrix form more than usual, it is because that form is less well known.* We believe that the matrix formulation must become familiar, as the teaching and practice of signal processing rely increasingly on computer systems like MATLAB.

Our goal is to understand filters, through impulse responses and frequency responses and transfer functions. If you know signal processing as in the Oppenheim–Schafer text, go past the first sections. If you are learning the whole subject from scratch, these sections can help. Do not hesitate to use this chapter for reference, as you reach Chapter 4 and the heart of the book.

We begin with signals.

2.1 Signals, Samples, and Time-invariance

A discrete-time signal is a sequence of numbers. The sequence could be finite or infinite. Most signals in this book are *doubly infinite*; the index n goes from $-\infty$ to $+\infty$. Time has no start and no finish. The signals look like

$$x = (\ldots, x_{-1}, x_0, x_1, x_2, \ldots) \quad \text{or} \quad x = \begin{bmatrix} \cdots \\ x(-1) \\ x(0) \\ x(1) \\ x(2) \\ \cdots \end{bmatrix}.$$

Those components are real or complex numbers (usually real). One particular signal is of tremendous value. It is the *unit impulse* $x = \delta$:

$$\delta = (\ldots, 0, 0, 1, 0, 0, \ldots) \quad \text{has components} \quad \delta(n) = \begin{cases} 0, & n \neq 0 \\ 1, & n = 0. \end{cases} \tag{2.1}$$

The continuous-time analogue is the "delta function" $\delta(t)$ — also called a Dirac impulse. In one case n is an integer, in the other t is a real number. The standard notations are $n \in \mathbf{Z}$ and $t \in \mathbf{R}$.

Together with the special vector δ goes the delayed impulse $S\delta$, where the unit component appears one sample later at $n = 1$. The whole vector is shifted by one time step. The symbol S stands for the *shift* or *delay* that has this effect on the vector δ:

$$S\delta = (\ldots, 0, 0, 0, 1, 0, \ldots) \quad \text{has components} \quad \delta(n-1) = \begin{cases} 0, & n \neq 1 \\ 1, & n = 1. \end{cases}$$

It is worth emphasizing that a shift *to the right* (a delay) produces the *minus sign* in the expression $n - 1$. It is the same in continuous time. The graph of $f(t)$, when it is shifted one unit to the right, is the graph of $f(t-1)$. The delayed function at $t = 1$ equals the original function at $t = 0$.

When the components are shifted to the left, the impulse comes sooner (at $n = -1$). This shift is an *advance* instead of a delay. The symbol is S^{-1}. The operator "S inverse" has an effect opposite to S:

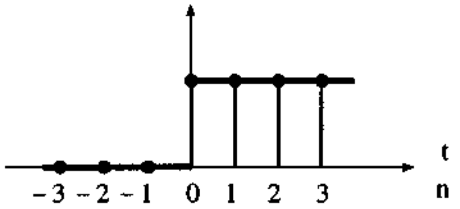
$$S^{-1}\delta = (\ldots, 0, 1, 0, 0, 0, \ldots) \quad \text{has components} \quad \delta(n+1) = \begin{cases} 0, & n \neq -1 \\ 1, & n = -1. \end{cases}$$

Of course $S^{-1}S\delta = \delta$. The impulse is delayed by S and then advanced by S^{-1}. Also $SS^{-1}\delta = \delta$. The operator S^{-1} is a "two-sided inverse" of S.

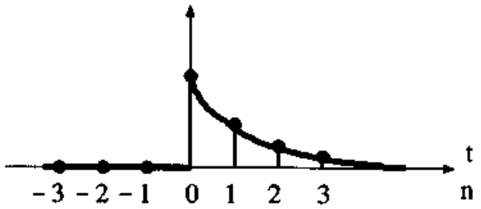
The vector δ could be defined, and Dirac's delta function should be defined, by what happens for inner products. The inner product (dot product) with any vector $x(n)$ or any continuous function $x(t)$ picks out $x(0)$:

$$x^T\delta = \sum_{-\infty}^{\infty} x(n)\delta(n) = x(0) \quad \text{and} \quad (x(t), \delta(t)) = \int_{-\infty}^{\infty} x(t)\delta(t)\,dt = x(0).$$

The number $x(0)$ is a sample of the function $x(t)$. Many discrete signals come from sampling continuous signals. This *analog to digital* (A/D) conversion is a central part of communications technology. We sketch two continous-time signals $x(t)$, a step and an exponential, and their discrete-time samples $x(n)$.



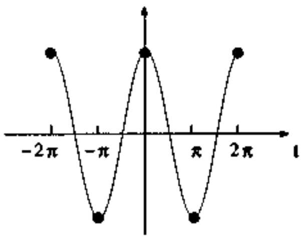
Unit Step



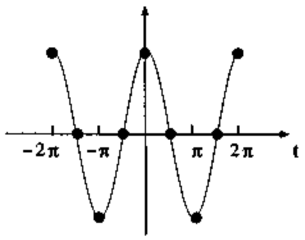
One-sided exponential

The samples of $\cos\omega t$ are of special importance. The continuous signal has frequency ω (often normalized as $f = \frac{\omega}{2\pi}$). Everything depends on the sampling period T and the sampling rate $f_s = \frac{1}{T}$. The sampling may be fast enough to catch the oscillations in $\cos\omega t$, or it may be too slow. We may catch the oscillations in $\cos\omega t$, or miss some. The borderline is the ***Nyquist rate***. The sampling rate is exactly the Nyquist rate when $f_s = 2f$ and $\frac{1}{T} = \frac{\omega}{\pi}$ and $\omega T = \pi$. This is the rate (two samples per oscillation) in the first figure. Those samples have the fastest oscillation that a discrete vector can achieve: $x(n) = (-1)^n$.

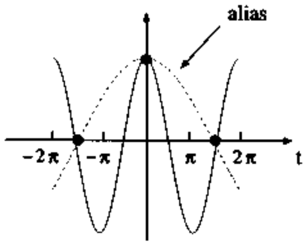
To repeat: The Nyquist rate gives the highest possible frequency $\omega T = \pi$.



Sampled at Nyquist rate



Twice the Nyquist rate



Slower than Nyquist rate

The second sampling rate is *faster than Nyquist.* So the digital frequency ωT is less than π. In this figure, the sampling frequency f_s is twice the Nyquist rate, four samples (*four bullets*) per oscillation. The sampled signal is $x(n) = \cos\omega nT = \cos\frac{n\pi}{2}$. Therefore the samples are $1, 0, -1, 0, 1, 0, \ldots$.

The third sampling rate is *slower than Nyquist.* The sample after $\cos 0$ is $\cos\frac{3\pi}{2}$. The sampling rate is $\frac{2}{3}$ of the Nyquist rate ($\frac{1}{T} = \frac{2}{3}\frac{\omega}{\pi}$). At this slow rate, the samples are the same $1, 0, -1, 0, 1, 0, \ldots$ as in the second figure! We cannot tell whether the true frequency of the continuous signal is ω (as drawn with solid line) or a slower frequency $\frac{\omega}{3}$ (as drawn by dotted line). ***This is aliasing***. The slow frequency $\frac{\omega}{3}$ is an alias for the true frequency ω, because the discrete samples at this rate will be exactly the same.

An extreme case of slow sampling is when $\omega T = 2\pi$. All the samples are $\cos\omega nT = \cos 2\pi n = 1$. Every sample is at the top of the wave. The digital frequency 2π looks identical to frequency zero (which is the alias of 2π).

When the continuous signal is a combination of many frequencies ω (or $f = \frac{\omega}{2\pi}$), the largest one ω_{max} sets the Nyquist rate $2f_{max}$. The corresponding Nyquist period T has $\omega_{max}T = \pi$. As long as the sampling period is smaller than this T, the sampling rate is faster than the Nyquist rate. Then there is no aliasing (by a lower frequency than the true frequency). *The continuous signal $x(t)$ can be recovered from its samples $x(n)$.* The Shannon sampling formula, to achieve that recovery, is in Section 2.2.

Impulses and Delays in Three Domains

Impulses are the building blocks for all signals. Delays are the building blocks for all filters. We will present signals and filters in three ways — with time variable n, and frequency variable ω, and complex variable z.

$$\begin{array}{lrcl} \text{Signal in the time domain} & x(n) & = & (\ldots, x(-1), x(0), x(1), \ldots) \\ \text{Signal in the frequency domain} & X(e^{j\omega}) & = & \sum x(n)e^{-j\omega n} \ \text{(standard)} \\ & X(\omega) & = & \sum x(n)e^{-i\omega n} \ \text{(reduced)} \\ \text{Signal in the } z\text{-domain} & X(z) & = & \sum x(n)z^{-n}. \end{array}$$

The standard notation and reduced notation were compared in Section 1.1. We use both! Sometimes the standard notation is clearer; it allows direct replacement of $e^{j\omega}$ by z. Sometimes the reduced notation is simpler, as in $Y(\omega) = H(\omega)X(\omega)$.

The impulse $\delta = (\ldots, 0, 0, 1, 0, 0, \ldots)$ becomes the constant function "1" in the frequency domain and z-domain. The only nonzero component $\delta(0) = 1$ is in the constant term. The delayed impulse $y = (\ldots, 0, 0, 0, 1, 0, \ldots)$ looks more interesting. In the other domains this $y(n) = \delta(n-1)$ is $Y(e^{j\omega}) = e^{-j\omega}$ and $Y(z) = z^{-1}$.

Note that $y(1) = 1$ multiplies the *negative power* of z, by the signal processing convention. If the impulse is advanced instead of delayed, the nonzero occurs at $n = -1$. The transform is z instead of z^{-1}. This signal is no longer causal. The advance operator S^{-1} is not a causal filter.

Now we define filters in general and study delays in particular.

A ***digital filter*** is a combination $\boldsymbol{H} = \sum \boldsymbol{h}(n)\boldsymbol{S}^n$ of delays S and advances S^{-1}.

The filter is completely determined by its coefficients $\boldsymbol{h}(n)$. When this sequence is finite, we have an "FIR filter". When $\boldsymbol{h}(n) = 0$ for negative n, we have a "causal filter". Our greatest interest is in causal FIR filters like

$$\boldsymbol{H} = \frac{1+\sqrt{3}}{8}\boldsymbol{I} + \frac{3+\sqrt{3}}{8}\boldsymbol{S} + \frac{3-\sqrt{3}}{8}\boldsymbol{S}^2 + \frac{1-\sqrt{3}}{8}\boldsymbol{S}^3 \quad \text{(Daubechies } D_4 \text{ filter).}$$

Suppose this filter acts on the impulse δ. The output is a combination of δ and its delays $\boldsymbol{S}\delta$ and $\boldsymbol{S}^2\delta$ and $\boldsymbol{S}^3\delta$:

$$\boldsymbol{H}\delta = (\ldots, \frac{1+\sqrt{3}}{8}, \frac{3+\sqrt{3}}{8}, \frac{3-\sqrt{3}}{8}, \frac{1-\sqrt{3}}{8}, 0, \ldots)$$

This is the "impulse response." It is equal to the vector $\boldsymbol{h}$ of filter coefficients:

The ***impulse response*** (causal and FIR) is $\boldsymbol{h} = (\boldsymbol{h}(0), \boldsymbol{h}(1), \ldots, \boldsymbol{h}(N))$.

You see why $\boldsymbol{H} = \sum \boldsymbol{h}(n)\boldsymbol{S}^n$ acting on δ produces this output $\boldsymbol{h}$. Each term $\boldsymbol{h}(n)\boldsymbol{S}^n$ produces one response $\boldsymbol{h}(n)$ at time n.

As displayed, the filter is FIR and causal with $N+1$ "taps". The filter length is even when N is odd! Some authors end at $h(N-1)$, so the length is N — but then the power $z^{-(N-1)}$ enters into a large number of formulas. We prefer to have sums from 0 to N, and scaling functions and wavelets on the interval $0 \le t < N$, and z^{-N} in all those formulas. The Daubechies filter has length 4 because $N = 3$.

The delay S takes $x = (\dots, x(0), x(1), \dots)$ into $y = (\dots, x(-1), x(0), \dots)$. Every linear operator like S is represented by a matrix. Since x and $y = Sx$ have infinitely many components, the shift matrix S has infinitely many rows and columns. This causal operator becomes a ***lower triangular*** matrix:

$$Sx = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & 0 & 0 & 0 & \cdot \\ \cdot & 1 & 0 & 0 & 0 & \cdot \\ \cdot & 0 & 1 & 0 & 0 & \cdot \\ \cdot & 0 & 0 & 1 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ x(-1) \\ x(0) \\ x(1) \\ x(2) \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ x(-2) \\ x(-1) \\ x(0) \\ x(1) \\ \cdot \end{bmatrix}.$$

The only nonzero coefficient for this filter is $h(1) = 1$. This coefficient goes along diagonal ***one***. In general $h(n)$ goes on diagonal n.

What does the delay do in the z-domain? The input $X(z)$ and output $Y(z)$ are

$$X(z) = \cdots + x(0) + x(1)z^{-1} + \cdots \quad \text{and} \quad Y(z) = \cdots + x(0)z^{-1} + x(1)z^{-2} + \cdots$$

The delay has multiplied $X(z)$ by z^{-1} to produce $Y(z)$. This transfer function z^{-1} is exactly the z-transform of the vector $(\dots, 0, 1, 0, 0, \dots)$ of filter coefficients. The pattern is always $Y(z) = H(z)X(z)$. Here is the special result for this particular filter, a delay $H = S$:

$$X(e^{j\omega}) \text{ is multiplied by } e^{-j\omega} \text{ and } X(z) \text{ is multiplied by } z^{-1}.$$

Time-invariant Filters

Our filters H are *linear*. This means in particular that "zero in produces zero out". If $x = 0$ then necessarily $y = 0$. The output from $2x$ is $2Hx$. The output from $x + z$ is $Hx + Hz$. This has an important consequence: *H is represented by a matrix.*

Our filters are also *time-invariant* (meaning shift-invariant). This leads to a special constant-diagonal property of the matrix:

$H(Sx) = S(Hx)$: A shift of the input produces a shift of the output.

Each column of H is a delay of the previous column.

Each diagonal of H is constant and the nth diagonal contains $h(n)$.

Those are different statements of time-invariance. They imply that H is a combination of shift operators: Every filter has the form $H = \sum h(n)S^n$. The filter $H = I + 4S + 3S^2$ has coefficients $h = (1, 4, 3)$. These are the entries down every column of the Toeplitz matrix.

$$\begin{array}{rcl}
\mathrm{H} = & \left[\begin{array}{cccccc}
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & 1 & 0 & 0 & 0 & \bullet \\
\bullet & 4 & 1 & 0 & 0 & \bullet \\
\bullet & 3 & 4 & 1 & 0 & \bullet \\
\bullet & 0 & 3 & 4 & 1 & \diagdown \\
\bullet & \bullet & \bullet & \diagdown & \diagdown & \diagdown
\end{array}\right] & \begin{array}{l} \\ \text{row } 0 \\ \text{row } 1 \\ \\ \\ -1 \\ \text{diagonal number} \end{array} \\
\text{column} & \begin{array}{cccccc} & 0 & 1 & & 2 & 1 \;\; 0 \end{array} &
\end{array}$$

Toeplitz matrix = constant-diagonal matrix with entries $\mathrm{H}_{ij} = \mathrm{h}(i-j)$.

The numbers $1, 4, 3$ also appear in every row—***but the order is reversed.*** This is a causal FIR filter. It is time-invariant, because $\boldsymbol{HS}$ or $\boldsymbol{SH}$ (they are the same!) is $S + 4S^2 + 3S^3$. All columns and all diagonals shift down by one, from the delay.

The difference between the row number i and the column number j is the diagonal number $k = i - j$. The entries of $\boldsymbol{H}$ depend only on k. This is a constant-diagonal matrix and a convolution matrix and a ***Toeplitz matrix***.

Matrix Multiplication and Vector Convolution

The jth column of $\boldsymbol{H}$ is $S^j\boldsymbol{h}$. We can compute $\boldsymbol{Hx}$ as a combination $\sum \boldsymbol{x}(j)(S^j\boldsymbol{h})$ of those columns. We can also compute $\sum \boldsymbol{h}(k)(S^k\boldsymbol{x})$. Best to see the nth component $\boldsymbol{Hx}(n)$:

$$\underset{\substack{\uparrow \\ n\text{th component of delayed } \boldsymbol{h}}}{\sum_j \boldsymbol{x}(j)\boldsymbol{h}(n-j)} \qquad = \qquad \underset{\substack{\uparrow \\ n\text{th component of delayed } \boldsymbol{x}}}{\sum_k \boldsymbol{h}(k)\boldsymbol{x}(n-k).}$$

The equality comes by changing j to $n-k$. The sums are over all integers, so the change is allowed. The finite sum to $j = N$ would not equal the sum to $k = N$, unless $\boldsymbol{h}$ has period N. (Then the N by N matrix $\boldsymbol{H}$ would be a ***circulant matrix***. This "wraparound" is the easiest way to deal with finite length signals, but not generally the best way.)

In both formulas for $\boldsymbol{Hx}(n)$, the indices add to n. The zeroth component is row zero of $\boldsymbol{H}$ times $\boldsymbol{x}$:

$$y(0) = \boldsymbol{Hx}(0) = \boldsymbol{h}(N)\boldsymbol{x}(-N) + \cdots + \boldsymbol{h}(1)\boldsymbol{x}(-1) + \boldsymbol{h}(0)\boldsymbol{x}(0). \tag{2.2}$$

Each pair of indices on the right adds to zero—which is the index on the left. The numbers $\boldsymbol{h}(N), \ldots, \boldsymbol{h}(0)$ are like a *moving window* that multiplies $\boldsymbol{x}$. The nth component of the output has indices adding to n:

$$\boldsymbol{Hx}(n) = \boldsymbol{h}(N)\boldsymbol{x}(n-N) + \cdots + \boldsymbol{h}(1)\boldsymbol{x}(n-1) + \boldsymbol{h}(0)\boldsymbol{x}(n). \tag{2.3}$$

This is the pattern that produces ***convolution***:

The output vector is $\boldsymbol{Hx} = \boldsymbol{h} * \boldsymbol{x} =$ *convolution of* $\boldsymbol{h}$ *with* $\boldsymbol{x}$.

Convolution Rule Indices automatically add when they are the exponents in a polynomial. Multiply $H(z)X(z)$ and their coefficients undergo a convolution:

$$H(z)X(z) = (\boldsymbol{h}(0)+\boldsymbol{h}(1)z^{-1}+\cdots+\boldsymbol{h}(N)z^{-N})(\cdots \boldsymbol{x}(-1)z+\boldsymbol{x}(0)+\boldsymbol{x}(1)z^{-1}+\cdots).$$

The coefficient of z^0 is $\boldsymbol{h}(0)\boldsymbol{x}(0)+\boldsymbol{h}(1)\boldsymbol{x}(-1)+\cdots+\boldsymbol{h}(N)\boldsymbol{x}(-N)$. This is $\boldsymbol{Hx}(0)$.

The coefficient of z^{-n} in the product $H(z)X(z)$ is the nth component of $\boldsymbol{Hx}$. Multiplying polynomials means collecting terms with the same exponent. This is convolution.

Example 2.1. The filter matrix has $\boldsymbol{h}(0)=1, \boldsymbol{h}(1)=4$, and $\boldsymbol{h}(2)=3$. Suppose the input has $\boldsymbol{x}(0)=1$ and $\boldsymbol{x}(1)=1$. Then we multiply polynomials or we take convolution of vectors:

$$\begin{array}{rrrr} & 3 & 4 & 1 \\ & & 1 & 1 \\ \hline & 3 & 4 & 1 \\ 3 & 4 & 1 & \\ \hline 3 & 7 & 5 & 1 \end{array} \qquad \begin{array}{rcl} (1+4z^{-1}+3z^{-2})(1+z^{-1}) & = & 1+5z^{-1}+7z^{-2}+3z^{-3} \\ (1,4,3) * (1,1) & = & (1,5,7,3). \end{array}$$

At $z=1$ this is 8 times 2 equals 16. At $z=-1$ we check 0 times 0 equals 0.

Example 2.2. Multiplying two filter matrices (Toeplitz matrices) is also a convolution. The product $\boldsymbol{FH}$ is another time-invariant filter, and its coefficients are in $\boldsymbol{f} * \boldsymbol{h}$. This is just like multiplying polynomials, with the shift S in place of the complex variable z^{-1}:

$$\boldsymbol{FH} = (\boldsymbol{I} + S)(\boldsymbol{I}+4S+3S^2) = \boldsymbol{I}+5S+7S^2+3S^3.$$

Note again that the order is not important: $\boldsymbol{FH}=\boldsymbol{HF}$. This will change when there are sampling operators $(\downarrow 2)$ and $(\uparrow 2)$ between the filters.

Example 2.3. The convolution of $(1, a, a^2, \ldots)$ with $(1, b, b^2, \ldots)$ is

$$(1+bz^{-1}+b^2z^{-2}+\cdots)(1+az^{-1}+a^2z^{-2}+\cdots) = 1+(a+b)z^{-1}+(a^2+ab+b^2)z^{-2}+\cdots$$

The power z^{-2} in the product comes from z^0 times z^{-2}, and from z^{-1} times z^{-1}, and from z^{-2} times z^0. The sum of exponents is -2, to give z^{-2} in the answer. This is the pattern for the indices k and $n-k$. Their sum is always n, to give $y(n)$ in the convolution.

The Inverse of a Time-invariant H

The filter $\boldsymbol{H}$ is ***invertible*** if and only if

$$\begin{array}{ll} H(\omega) \neq 0 & \text{for all frequencies } \omega \\ H(z) \neq 0 & \text{for all } |z| = |e^{j\omega}| = 1. \end{array} \tag{2.4}$$

Then $\boldsymbol{H}^{-1}$ is also a constant-diagonal matrix. Its frequency response is $1/H(\omega)$.

Invertibility is the first of many properties that become infinitely simpler by transforming convolution to $H(\omega)X(\omega)$. The inverse of multiplication is division! We recover $X(\omega)$ from $Y(\omega)/H(\omega)$. The requirement is $H(\omega) \neq 0$.

To emphasize: If we know a frequency ω_0 for which $H(\omega_0) = 0$, then we know an input $\boldsymbol{x}$ for which $\boldsymbol{Hx} = \mathbf{0}$. That input has the pure frequency ω_0. It is the vector with components $\boldsymbol{x}(n) = e^{-j\omega_0 n}$. The pure frequency is selectively killed by $H(\omega_0) = 0$. Then $H(\omega_0)X(\omega_0) = 0$ and $\boldsymbol{H}^{-1}$ fails.

A moving average with equal weights $\boldsymbol{h}(0) = \boldsymbol{h}(1) = \frac{1}{2}$ is not invertible. The frequency response $H(\omega) = \frac{1}{2}(1 + e^{-j\omega})$ is zero at $\omega = \pi$. The vector with components $\boldsymbol{x}(n) = e^{-j\pi n} = (-1)^n$ is exactly the vector that has $\boldsymbol{Hx} = \mathbf{0}$. By changing to two unequal weights *the system becomes invertible.*

Example 2.4. Suppose $\boldsymbol{h}(0) = 1$ and $\boldsymbol{h}(1) = -\beta$. The frequency response is $H(\omega) = 1 - \beta e^{-i\omega}$. If we select β smaller than one, then $1 \neq \beta e^{-i\omega}$. Thus $H(\omega) \neq 0$. The matrix $\boldsymbol{H}$ has 1 on the main diagonal and $-\beta$ on the diagonal below. To invert in the frequency domain, divide by $H(\omega)$. To invert in the time domain, practice with a 4 by 4 matrix:

$$\begin{bmatrix} 1 & & & \\ -\beta & 1 & & \\ & -\beta & 1 & \\ & & -\beta & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & & & \\ \beta & 1 & & \\ \beta^2 & \beta & 1 & \\ \beta^3 & \beta^2 & \beta & 1 \end{bmatrix}. \tag{2.5}$$

This suggests the correct diagonals $1, \beta, \beta^2, \ldots$ for the infinite matrix $\boldsymbol{H}^{-1}$. If $\boldsymbol{H}$ is $\boldsymbol{I} - \beta \boldsymbol{S}$, its inverse $\boldsymbol{I} + \beta \boldsymbol{S} + \beta^2 \boldsymbol{S}^2 + \cdots$ has the frequency response $1/H(\omega)$:

$$\frac{1}{1 - \beta e^{-i\omega}} = 1 + \beta e^{-i\omega} + (\beta e^{-i\omega})^2 + (\beta e^{-i\omega})^3 + \cdots$$

The most important of all infinite series (the *geometric series*) gives us this inverse: $\frac{1}{1-\beta} = 1 + \beta + \beta^2 + \cdots$. The sum is restricted to $|\beta| < 1$. Otherwise the series diverges.

Example 2.5. What if β is larger than 1? For a finite matrix we don't notice the difference. The 4 by 4 inverse above is still correct. But the infinite series has to be written *in powers of* $1/\beta$. *The inverse matrix changes from causal to anticausal.* Look first at the frequency response $1/H(\omega)$:

$$\frac{1}{1 - \beta e^{-i\omega}} = \frac{e^{i\omega}/\beta}{(e^{i\omega}/\beta) - 1} = -\frac{e^{i\omega}}{\beta} - \frac{e^{2i\omega}}{\beta^2} - \frac{e^{3i\omega}}{\beta^3} - \cdots. \tag{2.6}$$

This involves positive powers. We have *advances instead of delays* in the inverse.

The difference between $|\beta| < 1$ and $|\beta| > 1$ is the difference between a zero *inside* the unit circle and a zero *outside* that circle. $H(z) = 1 - \beta z^{-1}$ has its only zero at $z = \beta$. Here is the general rule for inverses of causal FIR systems:

No inverse when a zero is *on* the unit circle : $\boldsymbol{I} + \boldsymbol{S}$ has no inverse.

Causal inverse when all zeros are *inside* the unit circle: $1 - \beta z^{-1}$ has $|\beta| < 1$.

Anticausal inverse when all zeros are *outside* $|z| = 1$: $1 - \beta z^{-1}$ has $|\beta| > 1$.

A zero on the unit circle gives a particular frequency ω_0 at which $H = 0$. Then $1/H$ breaks down and the system has no inverse.

Problem Set 2.1

1. What matrix represents the inverse shift $y = S^{-1}x$? In the z-domain, the input is $X(z) = \cdots + x(0) + x(1)z^{-1} + \cdots$ and the output is $Y(z) =$ ___. The advance S^{-1} multiplies the z-transform by ___.

2. Express the filter $H = S+S^{-1}$ with coefficients $h(-1) = 1$ and $h(1) = 1$ in all three domains. Write the matrix H with two nonzero diagonals. Write the transfer function $H(z)$ with two terms. Write the frequency response $H(e^{j\omega})$ or $H(\omega)$, and check that this is the transform of the impulse response $h = H\delta$.

3. Show that $H=S+S^{-1}$ is not invertible in three ways. Find a nonzero input x such that $Hx = 0$. Find a frequency ω that has response $H(e^{j\omega}) = 0$. Find a number with $|z| = 1$ such that $H(z) = 0$.

4. What are the matrix H and coefficient vector h for the 3-term moving average $Hx(n) = \frac{1}{3}(x(n) + x(n-1) + x(n-2))$? This is not invertible. Find two vectors x for which $Hx = 0$. Find two numbers with $|z| = 1$ such that $H(z) = 0$. Find two frequencies such that $H(\omega) = 0$.

5. Express this 3-term moving average in the form $H = \sum h(n)S^n$. What is N? Find the output y when the input has $x(0) = x(1) = 1$. In the z-domain what are $X(z)$ and $Y(z)$, and how are they related?

6. For matrices show that $SS^{-1} = I$. What is the corresponding statement in the z-domain, about the transfer functions of S and S^{-1}?

7. Multiply the matrix S by itself. The product $H = S^2$ corresponds to what coefficient vector h and what transfer function $H(z)$?

8. Every filter $\sum h(n)S^n$ commutes with a delay: $\sum h(n)S^{n+1}$ is HS and also SH. Why does every filter commute with every other filter?

9. If the continuous-time signal is $x(t) = \cos t$, what is the period T that gives sampling exactly at the Nyquist rate? What samples $x(nT)$ do you get at this rate? What samples do you get from $x(t) = \sin t$?

10. If the sampling period is $T = 1$ and the continuous signal is $x(t) = e^{2\pi i t/5}$, describe the discrete signal $x(n)$. Is it periodic? Find two other frequencies ω such that $x(t) = e^{i\omega t}$ would give the same samples.

11. If the signal $x(t)$ has bandwidth 3 Khz, then the sampling rate must be at least ___ to avoid aliasing.

    Note that the sampling period is generally normalized to $T = 1$. Then the largest digital frequency is $\omega = \pi$. Our graphs of $H(\omega)$ do not extend beyond π.

12. Why is the downsampling operator $(\downarrow 2)x(n) = x(2n)$ not time-invariant? Give an example with $(\downarrow 2)Sx \neq S(\downarrow 2)x$.

13. When are these filters invertible? Which has a causal inverse? Which has an FIR inverse? Which is allpass with $|H(e^{j\omega})| = 1$?

$$\begin{aligned} H_1(z) &= (1-\alpha z^{-1})(1-\beta z^{-1}) \\ H_2(z) &= 1 + \beta z^{-1} + \beta^2 z^{-2} + \beta^3 z^{-3} + \cdots \\ H_3(z) &= (z-\beta)/(1-\beta z^{-1}) \\ H_4(z) &= 1 - \beta z^{-1} + z^{-2} \end{aligned}$$

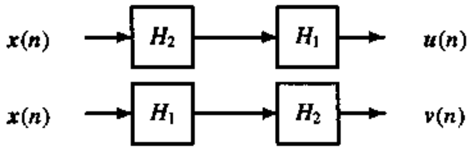
14. Determine the range of α and β for which the LTI system with impulse response $h(n) = \begin{cases} \alpha^n; & n \geq 0 \\ \beta^n; & n < 0 \end{cases}$ is stable. Find the output $y(n)$ when $x(n) = (-1)^n$.

15. Determine the impulse response for the cascade of the two LTI systems having impulse responses $h_1(n) = \left(\frac{1}{2}\right)^n u(n)$ and $h_2(n) = \left(\frac{1}{4}\right)^n u(n)$ using the convolution formula $h(n) = \sum h_1(k)h_2(n-k)$. Here $u(n)$ is the unit-step sequence.

16. Let H_1 be a system that throws away odd-indexed samples: $y(n) = x(2n)$. Is H_1 linear and time-invariant? H_1 is the downsampling block and its operation is discussed in Chapter 3.

17. Suppose H_2 inserts an extra zero between samples of the input: $y(n) = \begin{cases} x(n/2), & \text{even } n \\ 0, & \text{odd } n \end{cases}$.
Is H_2 a linear time-invariant system? H_2 is the upsampling block and its operation is discussed in Chapter 3.

18. Which cascade of downsampling and upsampling is time-invariant and what is its impulse response?



19. Give two examples of LTI systems, two examples of linear time-varying systems, and two examples of nonlinear systems.

2.2 Ideal Filters, Shannon Sampling, Sinc Wavelets

The word *filter* suggests that H selects a band of frequencies. It rejects another band. For ω in the ***passband***, the frequency response is near to $H(\omega) = 1$. For ω in the ***stopband***, the response is near to $H(\omega) = 0$. Any realizable non-ideal filter has a ***transition band*** in between, where $H(\omega)$ changes from pass to stop (from near 1 to near 0).

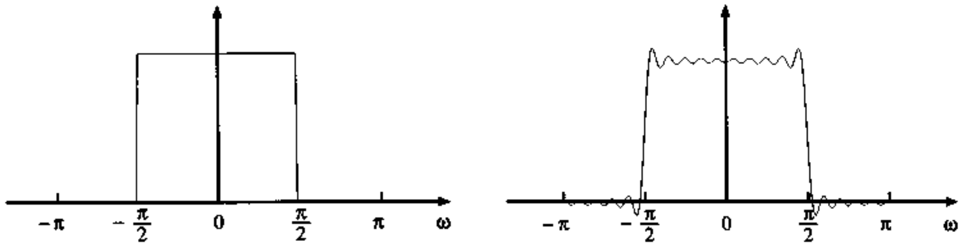


Figure 2.1: Ideal lowpass filter and best mean-square approximation with 20 terms.

We begin with an ideal filter, which has no transition band. Its responses are exactly $H(\omega) = 1$ and $H(\omega) = 0$. This is often called a *brick wall filter*, because of the step function in its graph. The response from an ideal lowpass filter is shown in Figure 2.1. This is a halfband filter, with sharp cutoff at $\omega = \frac{\pi}{2}$. The response $H(\omega) = \sum \boldsymbol{h}(k)e^{-ik\omega}$ is 2π-periodic, and the ideal response is zero in the high frequency band from $\frac{\pi}{2}$ to π:

$$\textbf{\textit{Ideal lowpass}} \quad H(\omega) = \sum_{-\infty}^{\infty} \boldsymbol{h}(k)e^{-ik\omega} = \begin{cases} 1, & 0 \le |\omega| < \frac{\pi}{2} \\ 0, & \frac{\pi}{2} \le |\omega| < \pi. \end{cases} \tag{2.7}$$

What filter coefficients $\boldsymbol{h}(k)$ produce this response? *Multiply the equation for $H(\omega)$ by $e^{in\omega}$ and integrate from $-\pi$ to π:*

$$\int_{-\pi}^{\pi} H(\omega)e^{in\omega}\,d\omega = \int_{-\pi}^{\pi}\Big(\sum_{k=-\infty}^{\infty} \boldsymbol{h}(k)e^{-ik\omega}\Big)e^{in\omega}\,d\omega. \tag{2.8}$$

On the right side, there is an integral of $e^{-ik\omega}e^{in\omega}$ for each k. The great property of complex exponentials is that this integral is zero except when $k = n$:

$$\int_{-\pi}^{\pi} e^{-ik\omega}e^{in\omega}\,d\omega = \left[\frac{e^{i(n-k)\omega}}{i(n-k)}\right]_{-\pi}^{\pi} = 0 \quad \text{if} \quad k \neq n. \tag{2.9}$$

The function in brackets is periodic. It has the same value at $-\pi$ and π. After substituting those limits, the definite integral is zero. ***The complex exponentials are orthogonal.***

Now equation (2.8) has only one term on the right, from $k = n$. Integrating this constant from $-\pi$ to π gives the result $2\pi \boldsymbol{h}(n)$. This equals the left side:

$$\int_{-\pi}^{\pi} H(\omega)e^{in\omega}\,d\omega = 2\pi \boldsymbol{h}(n). \tag{2.10}$$

The brick wall filter has $H(\omega) = 1$ on the part $|\omega| < \frac{\pi}{2}$:

$$2\pi \boldsymbol{h}(n) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{in\omega}\,d\omega = \left[\frac{e^{in\omega}}{in}\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \frac{2}{n}\sin\frac{\pi n}{2}. \tag{2.11}$$

The coefficients in the ideal lowpass filter are samples of a sinc function:

$$\boldsymbol{h}(n) = \frac{\sin\frac{\pi n}{2}}{\pi n} = \begin{cases} \frac{1}{2}, & n = 0 \\ \pm\frac{1}{\pi n}, & n \text{ odd} \\ 0, & n \text{ even}, n \neq 0. \end{cases} \tag{2.12}$$

The halfband cutoff has produced a halfband filter! The coefficient $\boldsymbol{h}(0) = \frac{1}{2}$ is the "DC term" = average value of $H(\omega)$. *All other even-numbered coefficients are* $\boldsymbol{h}(n) = 0$. When $H(\omega)$ is antisymmetric around the halfband frequency $\omega = \frac{\pi}{2}$, the filter is always halfband.

For odd n, the numbers $\boldsymbol{h}(n)$ alternate sign and decay slowly:

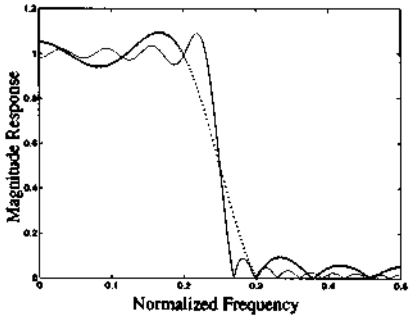
$$\boldsymbol{h}(1) = \boldsymbol{h}(-1) = -\frac{1}{\pi}, \quad \boldsymbol{h}(3) = \boldsymbol{h}(-3) = \frac{1}{3\pi}, \quad \boldsymbol{h}(5) = \boldsymbol{h}(-5) = -\frac{1}{5\pi}, \ldots$$

The series that adds up to the brick wall (= square wave = ideal lowpass response) is

$$H(\omega) = \frac{1}{2} - \frac{e^{i\omega} + e^{-i\omega}}{\pi} + \frac{e^{i3\omega} + e^{-i3\omega}}{3\pi} - \frac{e^{i5\omega} + e^{-i5\omega}}{5\pi} + \cdots. \tag{2.13}$$

At $\omega = \frac{\pi}{2}$, the only nonzero term is halfway down the brick wall: $H\left(\frac{\pi}{2}\right) = \frac{1}{2}$. Most important is the behavior *close to this jump at* $\omega = \frac{\pi}{2}$, as shown in Figure 2.1 and below. Suppose we chop off the series after N terms:

> The ripple at $\omega = \frac{\pi}{2}$ gets narrower as $N \to \infty$ but its height approaches a constant (about 0.09). This is the ***Gibbs phenomenon***.



This Gibbs phenomenon can be a disaster numerically. The ripple represents error. It is expensive to take a large number of terms and impossible to take all terms. A finite N gives the best approximation in the *mean square sense* — but the tall ripple remains. This "sidelobe" shows up as an echo in audio filtering and as a ghost in image processing. Practical design turns toward ***equiripple filters***, which have many ripples of equal height. This design minimizes the maximum ripple height instead of the total ripple energy.

Note however that equiripple filters do not behave well in iteration. *They do not lead to good wavelets.*

Minimax filter design is implemented by the Parks-McClellan algorithm, which computes best approximations to ideal filters. Those filters have a passband, a stopband, and a "don't care" transition band. The ripple heights (maximum errors) decay exponentially with the filter length N. If the acceptable error is specified, there is a formula for N (see *equiripple* in the Glossary). An alternative is *eigenfilter design*. For many problems this allows a simple mean square calculation, but without the big sidelobe from the Gibbs phenomenon.

Historical note. It is surprising to read the original paper by Gibbs. He completely missed the Gibbs phenomenon. His correction published later was even shorter — about three important lines. This correction must have the highest signal to noise ratio in the history of science.

Fourier's Series by J. Willard Gibbs [*Nature*, vol. LIX, p. 200, December 29, 1898.]

> . . . Let us write $f_n(x)$ for the sum of the first n terms of the series
>
> $$\sin x - \tfrac{1}{2}\sin 2x + \tfrac{1}{2}\sin 3x - \tfrac{1}{4}\sin 4x + \text{etc.}$$
>
> As n increases without limit, the curve defined by $y = 2f_n(x)$ approaches a limiting form, which may be thus described. Let a point move from the origin in a straight line at an angle of 45° with the axis of X to the point (π, π), thence vertically in a straight line to the point $(\pi, -\pi)$, thence obliquely in a straight line to the point $(3\pi, \pi)$. The broken line thus described (continued indefinitely forwards and backwards) is the limiting form of the curve as the number of terms increases indefinitely. . . .

Correction [in *Nature*, vol. LIX, p. 606, April 27, 1899.]

> I should like to correct a careless error which I made in describing the limiting form of the family of curves represented in the equation
>
> $$y = 2\left(\sin x - \tfrac{1}{2}\sin 2x \cdots \pm \tfrac{1}{n}\sin nx\right) \qquad (2.14)$$
>
> as a zigzag line consisting of alternate inclined and vertical portions. The inclined portions were correctly given, but the vertical portions, which are bisected by the axis of X, extend beyond the points where they meet the inclined portions, their total lengths being expressed by four times the definite integral $\int_0^\pi \frac{\sin u}{u}\,du$. . . . But this limiting form of the graphs of the functions expressed by the sum is different from the graph of the function expressed by the limit of that sum.
>
> I think this distinction important, for (with exception of what relates to my unfortunate blunder described above) whatever differences of opinion have been expressed on this subject seem due, for the most part, to the fact that some writers have had in mind the *limit of the graphs*, and others the *graph of the limit* of the sum.

The Gibbs phenomenon means that convergence to a brick wall is *not uniform*, as N increases. The coefficients $h(n)$ approach zero but the sum of absolute values $1 + \frac{1}{3} + \frac{1}{5} + \cdots$

is infinite. These are multiplied by $\pm \sin n\omega$, so the actual series for $H(\omega)$ does not blow up. But the slow $1/n$ decay prevents uniform convergence and allows the large sidelobe. This ripple always appears in the Fourier series near a jump discontinuity.

Ideal Filter with Downsampling

In the time domain, $\boldsymbol{h}(n)$ is on the nth diagonal of the filter matrix $\boldsymbol{H}$. Writing a and b in place of $\boldsymbol{h}(1) = \boldsymbol{h}(-1)$ and $\boldsymbol{h}(3) = \boldsymbol{h}(-3)$, three rows are

$$\boldsymbol{H} = \begin{array}{ccccccccccccc} \cdots & 0 & b & 0 & a & 0.5 & a & 0 & b & 0 & \cdots & & \\ & \cdots & 0 & b & 0 & a & 0.5 & a & 0 & b & 0 & \cdots & \\ & & \cdots & 0 & b & 0 & a & 0.5 & a & 0 & b & 0 & \cdots \end{array}$$

Those rows are not orthogonal! The dot product of the first two rows is a. Only an allpass filter has orthonormal rows (and columns). Then $|H(\omega)| = 1$ for all ω.

What is fundamental for this book is that *the even-numbered rows of $\boldsymbol{H}$ are orthogonal.* When the downsampling operator $(\downarrow 2)$ removes half of the rows, this leaves a ***double shift*** in the remaining rows — the rows of $(\downarrow 2)\boldsymbol{H}$:

$$(\downarrow 2)\boldsymbol{H} = \begin{array}{ccccccccccccccc} \cdots & 0 & b & 0 & a & 0.5 & a & 0 & b & 0 & \cdots & & & & \\ & & \cdots & 0 & b & 0 & a & 0.5 & a & 0 & b & 0 & \cdots & & \\ & & & & \cdots & 0 & b & 0 & a & 0.5 & a & 0 & b & 0 & \cdots \end{array}$$

Orthogonality is not so clear in this time domain. Moving into the frequency domain, the double shift is a multiplication by $e^{-i2\omega}$ and row 0 of $\boldsymbol{H}$ is orthogonal to row 2:

$$(\text{row } 0) \cdot (\text{row } 2) \quad = \quad \int_{-\pi}^{\pi} H(\omega)\overline{e^{-i2\omega}H(\omega)}\, d\omega \quad = \quad \int_{-\pi/2}^{\pi/2} e^{i2\omega}\, d\omega \quad = \quad 0.$$

Similarly row 0 is orthogonal to row 4, because the integral of $e^{i4\omega}$ is zero. This integral is over the half-period where $H(\omega) = 1$. The integral of $e^{i\omega}$ is not zero over this half-period, and row 0 of $\boldsymbol{H}$ is not orthogonal to row 1.

The Ideal Highpass Filter

Haar's lowpass filter has coefficients $\frac{1}{2}$ and $\frac{1}{2}$, where the highpass filter has $\frac{1}{2}$ and $-\frac{1}{2}$. Those are clearly orthogonal. Now the ideal lowpass filter has infinitely many coefficients. We want to construct a highpass filter $\boldsymbol{H}_1$, so that the rows of $(\downarrow 2)\boldsymbol{H}_1$ will be orthogonal to the rows of $(\downarrow 2)\boldsymbol{H}$.

It is easy to make $H_1(\omega)$ orthogonal to the ideal lowpass $H(\omega)$. We set $H_1 = 1$ in the intervals where $H = 0$:

$$\text{The ideal } H_1(\omega) \text{ can be } \begin{cases} 0 & \text{when} \quad 0 \le |\omega| < \frac{\pi}{2} \\ 1 & \text{when} \quad \frac{\pi}{2} \le |\omega| < \pi. \end{cases}$$

Shifted by π, the ideal $H(\omega)$ produces $H_1(\omega)$. In the time domain, that shift by π reverses the signs of the *odd-numbered* components (because $\boldsymbol{h}(k)$ changes to $\boldsymbol{h}(k)e^{ik\pi}$):

$$\boldsymbol{h}_1(0) = \frac{1}{2}, \quad \boldsymbol{h}_1(1) = \boldsymbol{h}_1(-1) = +\frac{1}{\pi}, \quad \boldsymbol{h}_1(3) = \boldsymbol{h}_1(-3) = -\frac{1}{3\pi}, \ \ldots$$

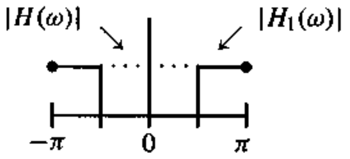


Figure 2.2: Ideal magnitude frequency responses: low $|H(\omega)|$ and high $|H_1(\omega)|$.

The graphs of $|H(\omega)|$ and $|H_1(\omega)|$ are in Figure 2.2. We explain below why absolute values suddenly appeared. Whatever the phase, orthogonality is sure because of no overlap.

We note that $|H(\omega)|+|H_1(\omega)|=1$. This is not the identity that really matters. It is the sum of *squares* that applies not only in this case but in all orthogonal filter banks:

$$|H(\omega)|^2+|H(\omega+\pi)|^2\equiv 1. \tag{2.15}$$

For the Haar example this is the identity $\cos^2\frac{\omega}{2}+\sin^2\frac{\omega}{2}=1$. The ideal case is deceptive because $1=1^2$ and $0=0^2$. The orthogonality requirement (2.15) will be established in Section 5.2.

The Alternating Flip (with Odd Shift)

There is an unusual point about the step from $\boldsymbol{H}$ to $\boldsymbol{H}_1$. In the time domain this usually comes from three operations on the coefficients $\boldsymbol{h}(n)$: *reverse the order, alternate the signs, and shift by 1* (or any odd N). This takes the lowpass coefficients $\boldsymbol{h}(n)$ into an orthogonal highpass sequence:

$$\textbf{\textit{Alternating flip}} \quad \boldsymbol{h}_1(n)=(-1)^n\boldsymbol{h}(N-n). \tag{2.16}$$

For a finite sequence $\boldsymbol{h}(0),\boldsymbol{h}(1),\ldots,\boldsymbol{h}(N)$ — assuming N is odd! — you immediately see the flip in the next figure and the orthogonality between rows:

low	$\boldsymbol{h}(0)$	$\boldsymbol{h}(1)$	$\cdots$	$\boldsymbol{h}(N-1)$	$\boldsymbol{h}(N)$
high	$\boldsymbol{h}(N)$	$-\boldsymbol{h}(N-1)$	$\cdots$	$\boldsymbol{h}(1)$	$-\boldsymbol{h}(0)$

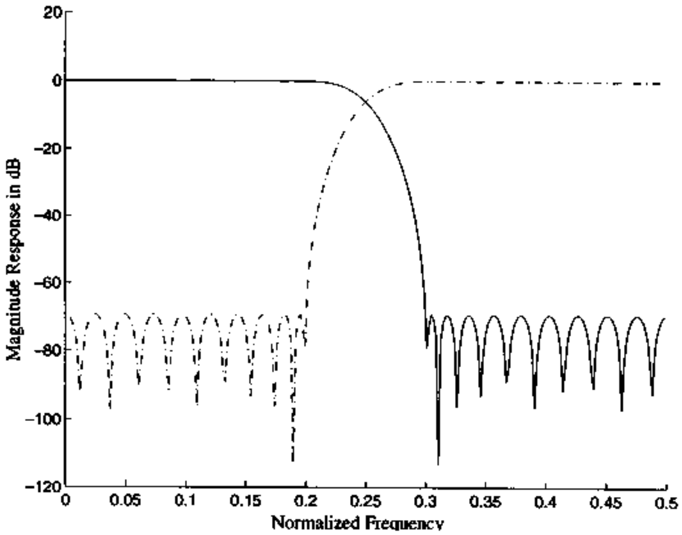
Important: There is also orthogonality of double shifts, as in

shift low by 2	$\boldsymbol{h}(2)$	$\boldsymbol{h}(3)$	$\cdots$	$\boldsymbol{h}(N-1)$	$\boldsymbol{h}(N)$
high	$\boldsymbol{h}(N)$	$-\boldsymbol{h}(N-1)$	$\cdots$	$\boldsymbol{h}(3)$	$-\boldsymbol{h}(2)$

$\boldsymbol{h}(2)\boldsymbol{h}(N)$ cancels $-\boldsymbol{h}(N)\boldsymbol{h}(2)$. This happens for all double shifts. It does not usually happen for single shifts! A single shift of Haar to $0,\frac{1}{2},\frac{1}{2}$ is not orthogonal to the highpass $\frac{1}{2},-\frac{1}{2},0$. Double shifts are all we care about, because it is double-shifted rows in $(\downarrow 2)\boldsymbol{H}$ that are orthogonal — and now the highpass rows in $(\downarrow 2)\boldsymbol{H}_1$ complete the filter bank.

The alternating flip is the key to orthogonality. In the frequency domain it has three steps: *Multiply by $e^{i\omega}$ to shift, take complex conjugates to flip, shift by π to alternate signs*:

$$H_1(\omega)=e^{-i\omega}\overline{H(\omega+\pi)}. \tag{2.17}$$



Theorem 2.1 *The alternating flip makes the rows of* $(\downarrow 2)\boldsymbol{H}$ *orthogonal to the rows of* $(\downarrow 2)\boldsymbol{H}_1$.

This was verified by eye, in the low and high rows above. You can verify it again for the ideal filters with infinitely many $\boldsymbol{h}$'s:

row 0 of $\boldsymbol{H}$	$\cdots$	b	0	a	0.5	a	0	b	0	$\cdots$
row 0 of $\boldsymbol{H}_1$	$\cdots$	0	$-b$	0	$-a$	0.5	$-a$	0	$-b$	$\cdots$

In the frequency domain we will see orthogonality in another form:

$$H(\omega)\overline{H_1(\omega)} + H(\omega+\pi)\overline{H_1(\omega+\pi)} = H(\omega)e^{i\omega}H(\omega+\pi) + H(\omega+\pi)e^{i(\omega+\pi)}H(\omega) \equiv 0. \tag{2.18}$$

All great, but for the ideal filters there is a very strange point. There was no odd shift! From the brick wall $H(\omega)$ on $[-\frac{\pi}{2}, \frac{\pi}{2}]$, we shifted by π to build the highpass brick wall. The wall is real, so conjugation has no effect. Still there should have been a phase shift from $e^{i\omega}$ and there wasn't.

Apparently row 0 of $\boldsymbol{H}$ is orthogonal to $[-b \ \ 0 \ \ -a \ \ .5 \ \ -a \ \ 0 \ \ -b \ \ 0 \ \cdots]$ *without the shift*. This unusual point occurs because $H_1(\omega)$ does not overlap $H(\omega)$. We can give $H_1(\omega)$ any phase we desire. It was natural to make it real. It is more consistent to include the phase shift $e^{i\omega}$.

This oddity (actually it is a lack of oddity) will reappear below for ideal wavelets. The sinc wavelets should have a shift in time, from the phase factor $e^{i\omega}$ — but generally they are taken from the unshifted $\boldsymbol{H}_1$. Before turning to scaling functions and wavelets, we include a short discussion of the Sampling Theorem — to recover a band-limited function $x(t)$ from its samples. This famous theorem appears everywhere, so we focus on a particular aspect involving $(\downarrow 2)$.

Shannon (Down-)Sampling Theorem

Our main theme for filter banks is perfect reconstruction of all signals. With two filters this will be achieved in spite of downsampling. The Sampling Theorem restricts the input to a subspace of band-limited signals. Then *one* downsampled output is enough to recover the input.

The signal lies in the lower halfband $|\omega| < \frac{\pi}{2}$; no higher frequencies are allowed. In an ideal filter bank, nothing comes out of the highpass channel. Full information must be in $(\downarrow 2)\boldsymbol{H}\boldsymbol{x}$. For a half-range of *input frequencies*, we only need a half-range of *output samples.*

Suppose the output $(\downarrow 2)\boldsymbol{x}$ *is an impulse* $\boldsymbol{\delta} = (\ldots, 0, 1, 0, \ldots)$. *What was* $\boldsymbol{x}$? A first suggestion is $\boldsymbol{x} = \boldsymbol{\delta}$, since $(\downarrow 2)\boldsymbol{\delta}$ is equal to $\boldsymbol{\delta}$. But this is wrong — because $\boldsymbol{\delta}$ is not band-limited. The impulse has all frequencies in equal amounts.

The correct input to yield $(\downarrow 2)\boldsymbol{x} = \boldsymbol{\delta}$ has *a halfband of frequencies in equal amounts.* The graph of $X(\omega)$ is a square wave:

$$X(\omega) = \begin{cases} 2 & \text{for} \quad 0 \le |\omega| < \frac{\pi}{2} \\ 0 & \text{for} \quad \frac{\pi}{2} \le |\omega| < \pi. \end{cases} \tag{2.19}$$

Downsampling doubles every frequency, so $(\downarrow 2)\boldsymbol{x}$ has a full band of frequencies in equal amounts. It equals $\boldsymbol{\delta}$ as required.

What signal $\boldsymbol{x}$ has the transform $X(\omega)$ = square wave? The inverse transform is

$$\boldsymbol{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)e^{in\omega}\,d\omega = \frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{in\omega}\,d\omega = \frac{2}{n\pi}\sin\frac{n\pi}{2}. \tag{2.20}$$

The input before downsampling has been recovered as the **sinc vector**:

$$\boldsymbol{x}_{\text{sinc}}(n) = \frac{\sin\frac{n\pi}{2}}{\frac{n\pi}{2}} \quad \text{with the convention} \quad \boldsymbol{x}_{\text{sinc}}(0) = 1. \tag{2.21}$$

Downsampling gives $\boldsymbol{\delta}$ *because if* n *is even then* $\sin\frac{n\pi}{2} = 0$.

The recovery problem is now solved when $(\downarrow 2)\boldsymbol{x}$ is $\boldsymbol{\delta}$. The band-limited input was $\boldsymbol{x}_{\text{sinc}}$. But every output is a combination of impulses at different times:

$$(\downarrow 2)\boldsymbol{x} = (\ldots, \boldsymbol{x}(0), \boldsymbol{x}(2), \ldots) = \cdots + \boldsymbol{x}(0)\boldsymbol{\delta} + \boldsymbol{x}(2)S\boldsymbol{\delta} + \cdots$$

Delaying the input by 2 delays $(\downarrow 2)\boldsymbol{x}$ by 1. This leads us to the correct input:

Downsampling Theorem The halfband signal that produces $(\downarrow 2)\boldsymbol{x} = (\ldots, \boldsymbol{x}(0), \boldsymbol{x}(2), \ldots)$ is

$$\boldsymbol{x}(n) = \cdots + \boldsymbol{x}(0)\boldsymbol{x}_{\text{sinc}}(n) + \boldsymbol{x}(2)\boldsymbol{x}_{\text{sinc}}(n-2) + \cdots = \sum_{-\infty}^{\infty} \boldsymbol{x}(2k)\frac{\sin\left((n-2k)\frac{\pi}{2}\right)}{(n-2k)\frac{\pi}{2}}.$$

For even n all terms are zero, except $2k = n$ which yields $\boldsymbol{x}(n)$. The input signal $\boldsymbol{x}$ is halfband because $\boldsymbol{x}_{\text{sinc}}$ and its shifts are halfband.

By changing $2k$ to k we would get the ordinary Shannon Sampling Theorem.

Sinc Wavelets (Shannon Wavelets)

In the Haar example, the lowpass $\boldsymbol{H}$ is the averaging filter (coefficients $\frac{1}{2}$ and $\frac{1}{2}$). By iteration we reached a continuous-time box function. That function satisfies the dilation equation with coefficients 1 and 1. The box is the "scaling function," and there is a corresponding up-down Haar wavelet.

Now $\boldsymbol{H}$ is the ideal lowpass filter. *Its scaling function is the sinc function* $\phi(t) = \frac{\sin\pi t}{\pi t}$. Since the ideal filter is IIR, the steps in our discussion will go a little more steeply. After this page, our main theme is FIR filter banks.

Section 6.4 has an infinite product formula for the Fourier transform of $\phi(t)$:

$$\widehat{\phi}(\omega) = \prod_{j=1}^{\infty} H\left(\frac{\omega}{2^j}\right). \tag{2.22}$$

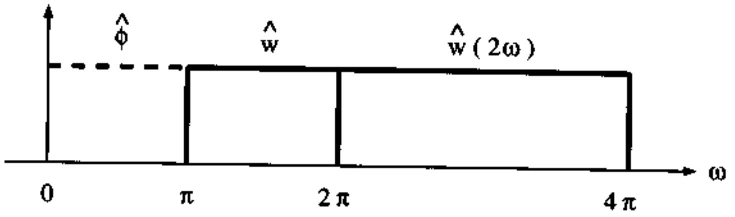
In the ideal case, every factor $H(\omega/2^j)$ is one for $|\omega| < \pi$. The jth factor is zero for $2^{j-1}\pi \le |\omega| < 2^j\pi$. The infinite product gives a box function for $\widehat{\phi}(\omega)$, stretching from $-\pi$ to π. (Note that $H(\omega)$ is 2π-periodic, but the infinite product $\widehat{\phi}(\omega)$ is not.) The inverse transform $\phi(t)$ of this box is a sinc function:

The ideal scaling function is $\phi(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega t}\, d\omega = \frac{\sin \pi t}{\pi t}$.

The wavelet $w(t) = \sum 2h_1(k)\phi(2t-k)$ comes from one application of the highpass filter (with downsampling) to $\phi(t)$. In the frequency domain this is

$$\widehat{w}(\omega) = H_1\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right) = \begin{cases} 1 & \text{for } \pi \le |\omega| < 2\pi \\ 0 & \text{otherwise.} \end{cases} \tag{2.23}$$

The ideal filter bank cuts the frequency band in half. The upper half of the band goes through the highpass filter (discrete time). In continuous time it is a combination of wavelets. The lower half of the frequency band goes through the lowpass filter (discrete time). In continuous time this half is a combination of scaling functions ϕ — ready to be split again into wavelets and scaling functions at the next finer scale. We have an *octave decomposition* = *logarithmic decomposition* = "*constant-Q decomposition*" of the line of frequencies:



Note 1 Meyer smoothed out this ideal picture to produce band-limited wavelets with fast decay (IIR of course). The key is to keep $\sum |\widehat{w}_{\text{Meyer}}(2^n\pi\omega)|^2 \equiv 1$. This can be done smoothly with overlap of nearest neighbors only [D]. All band-limited wavelets have two bumps ($\pm\omega$) like the Shannon and Meyer wavelets.

Problem Set 2.2

1. Show that the inverse transform of $\widehat{w}$ in (2.23) is the ***sinc wavelet*** $w(t) = 2\text{sinc}(2t) - \text{sinc}(t)$.
2. Find the shifted sinc wavelet by inverse transform when the factor $e^{i\omega}$ is included in $H_1(\omega)$. This is the odd shift that we normally need for orthogonality of w to ϕ.
3. What are the coefficients $h(n)$ for the ideal quarterband filter, with $H(\omega) = 1$ on $[-\frac{\pi}{4}, \frac{\pi}{4}]$? What scaling function $\widehat{\phi}(\omega)$ comes from the infinite product formula? Is $\phi(t)$ orthogonal to its translates?
4. ***(Important)*** Show that $H(\omega)$ has halfband symmetry (odd around its value at $\omega = \frac{\pi}{2}$) when $h(n)$ is a **halfband filter**. This means $(\downarrow 2)h = \delta$.
5. Let $h_{LP}(n)$ denote an ideal lowpass filter with cutoff frequency ω_c.

(a) Compute $H_{LP}(e^{j\omega})$ and normalized $\boldsymbol{h}_{LP}(n)$ such that $H_{LP}(e^{j0}) = 1$.

(b) The ideal highpass filter $\boldsymbol{h}_{LP}(n)$ can be designed by

$$\boldsymbol{h}_{HP}(n) = \begin{cases} 1 - \boldsymbol{h}_{LP}(n); & n = 0, \\ -\boldsymbol{h}_{LP}(n); & \text{otherwise.} \end{cases}$$

Find $H_{HP}(e^{j\omega})$.

(c) Design a highpass filter with cutoff frequency at $\pi - \omega_c$, using $\boldsymbol{h}_{LP}(n)$.

6. Let $H(z) = (1 - 2z^{-1} + 3z^{-2} - 3z^{-3} + 2z^{-4} - z^{-5})$. Compute $|H(e^{j\omega})|$ and the phase response $\phi(\omega)$ and the group delay $\phi'(\omega)$.

7. Let $H(z)$ be an FIR lowpass filter of length $(N+1)$. Define $G_1(z) = z^{-N}\, H(z^{-1})$, $G_2(z) = H(-z)$ and $G_3(z) = z^{-N} H(-z^{-1})$.

(a) What are $\boldsymbol{g}_1(n)$, $\boldsymbol{g}_2(n)$ and $\boldsymbol{g}_3(n)$ in terms of $\boldsymbol{h}(n)$?

(b) If $H(z)$ is an even-length symmetric filter, what is the symmetry or antisymmetry of $G_1(z)$, $G_2(z)$ and $G_3(z)$?

(c) If z_0 is a zero of $H(z)$, what are the corresponding zeros of $G_1(z)$, $G_2(z)$ and $G_3(z)$?

(e) What are the relations of $|H(e^{j\omega})|$, $|G_1(e^{j\omega})|$, $|G_2(e^{j\omega})|$, and $|G_3(e^{j\omega})|$?

8. Show that $G(z) = H(z)H(z^{-1})$ is a symmetric filter. What type of filter is $G(z)$ (lowpass, bandpass, highpass) if $H(z)$ is highpass? What is the (constant) phase of $G(z)$?

9. We have stated that $\delta_p = \delta_s$ in a halfband filter. Prove this.

2.3 Lowpass and Highpass Filter Design

The previous section dealt with ideal filters (necessarily IIR). This section deals with real FIR filters — often symmetric or antisymmetric (thus linear phase). We indicate the goals of filter design and we briefly discuss design methods.

For ideal brick walls, the transition from $H(\omega) = 1$ to $H(\omega) = 0$ happens instantly. For FIR filters this is not possible. It is important to see an actual magnitude response graph (Figure 2.3). In normal scale, we can observe details in the passband but not in the stopband. In logarithmic scale (dB scale, plotting $20\log_{10}|H|$), it is the other way around. The stopband details are visible in dB scale but the passband details are lost.

Before moving to good lowpass filters, we review two very short and rather poor filters. This gives us a chance to emphasize the four types of linear phase filters — odd and even length, symmetric and antisymmetric.

Example 2.6. The impulse response is $\boldsymbol{h} = \left(\frac{1}{2}, \frac{1}{2}\right)$. The frequency response is

$$H(\omega) = \tfrac{1}{2}\left(1 + e^{-i\omega}\right) = \left(\frac{e^{i\omega/2} + e^{-i\omega/2}}{2}\right)e^{-i\omega/2} = \left(\cos\frac{\omega}{2}\right)e^{-i\omega/2}. \tag{2.24}$$

In that last form you see the magnitude $|H(\omega)| = \cos\frac{\omega}{2}$ and the phase $\phi(\omega) = -\frac{\omega}{2}$. The magnitude is $\cos 0 = 1$ at zero frequency — this is a lowpass filter. The magnitude drops to zero at $\omega = \pi$. The *phase* $\phi(\omega)$ is the *angle* $-\omega/2$ in the polar form $re^{i\phi}$. *This phase function* $-\frac{\omega}{2}$ *is linear in* ω. The noninteger $1/2$ reflects the fact that the coefficients in $\boldsymbol{h}$ are symmetric about the "1/2" position.

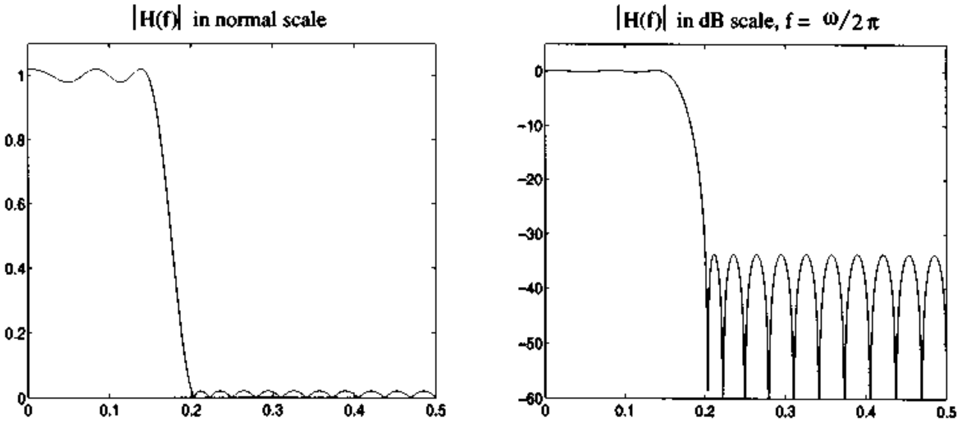


Figure 2.3: Magnitude response of a lowpass filter in normal and dB scale.

We will compute the phase $\phi(\omega)$ for other filters before analyzing its significance. Here we only mention: linear phase is a desirable property.

Example 2.7. By cascading the previous example we square its matrix $\boldsymbol{H}$ and we square its frequency response. The new filter coefficients are in

$$\begin{bmatrix} 0.5 & & \\ 0.5 & 0.5 & \\ 0 & 0.5 & 0.5 \\ \ddots & \ddots & \ddots \end{bmatrix}^2 = \begin{bmatrix} 0.25 & & \\ 0.50 & 0.25 & \\ 0.25 & 0.50 & 0.25 \\ \ddots & \ddots & \ddots \end{bmatrix} = \boldsymbol{H}_{\text{new}}.$$

The new impulse response is $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$. It is the convolution of $\left(\frac{1}{2}, \frac{1}{2}\right)$ with itself. Cascading filters means convolution of impulse responses (time domain). In the frequency domain we multiply responses $H(\omega)$:

$$H_{\text{new}}(\omega) = [H_{\text{old}}(\omega)]^2 = \tfrac{1}{4}(1 + e^{i\omega})^2 = \left(\cos^2 \frac{\omega}{2}\right)e^{-i\omega}. \tag{2.25}$$

This squares the magnitude and doubles the phase. The new phase $-\omega$ is still linear. It corresponds to a time shift of 1. The impulse response is a unit delay of a symmetric $\boldsymbol{h}$. $\boldsymbol{H}_{\text{new}}$ is a shift (= delay) times a symmetric matrix.

Every FIR matrix can be made causal, by sufficiently many delays. A symmetric matrix (not causal) corresponds to phase = zero because $H(\omega)$ is *real*:

$$H_{\text{sym}}(\omega) = \tfrac{1}{4}e^{i\omega} + \tfrac{1}{2} + \tfrac{1}{4}e^{-i\omega} = \tfrac{1}{2}(1 + \cos\omega). \tag{2.26}$$

Was this cascade desirable? Neither $\boldsymbol{H}$ or $\boldsymbol{H}_{\text{new}}$ is very impressive. The frequency responses are far from ideal. $H_{\text{new}}(\omega)$ has better attenuation in the stopband, because the cosine is squared. But it is also smaller in the passband — farther away from the ideal $H \equiv 1$.

By alternating signs in the lowpass coefficients they become *highpass*: $\boldsymbol{h}_1 = (0.5, -0.5)$ and $\boldsymbol{h}_1 * \boldsymbol{h}_1 = (0.25, -0.50, 0.25)$. These are still linear phase. The alternation of signs is a phase shift (a *modulation*) by π. The first is antisymmetric but the second is still symmetric.

It is useful to tabulate the four types of linear phase filters with real coefficients. N can be odd or even. The coefficients can satisfy

$$h(n) = h(N-n) \text{ for symmetric,} \quad h(n) = -h(N-n) \text{ for antisymmetric.}$$

The linear phase $\phi(\omega) = -\omega N/2$ and the (real) amplitude response $H_R(\omega)$ are seen in $H\left(e^{j\omega}\right) = ce^{-j\omega N/2}H_R(\omega)$. The table shows that with odd N, symmetry guarantees a zero at $\omega = \pi$ and antisymmetry guarantees a zero at $\omega = 0$. The responses $H_R(\omega)$ in the table have factors $\cos\frac{\omega}{2}$ and $\sin\frac{\omega}{2}$. Remember that the filter length (number of taps) is $N+1$. Here $c = 1$ for symmetric and $c = j$ for antisymmetric filters.

Type 1	Type 2	Type 3	Type 4
even $N = 2K$ symmetric	odd $N = 2K+1$ symmetric	even $N = 2K$ antisymmetric	odd $N = 2K+1$ antisymmetric
$H_R = \sum_0^K b_n \cos n\omega$	$\cos\frac{\omega}{2}\sum_0^K b_n \cos n\omega$ zero at $\omega = \pi$	$\sin\omega\sum_0^{K-1} b_n \cos n\omega$ zeros at $\omega = 0, \pi$	$\sin\frac{\omega}{2}\sum_0^K b_n \cos n\omega$ zero at $\omega = 0$

Now consider the ideal lowpass filter with cutoff frequency ω_c. This is a band-limited filter in the frequency domain, therefore its support in the time domain is infinite. It must be IIR. Its impulse response is $h_I(n) = \pi\frac{\sin(\omega_c n)}{\omega_c n}$. Since the time-support is infinite, one needs to approximate it by a finite impulse response. The sequence $h(n)$ becomes time-limited, therefore not band-limited. The magnitude response $|H\left(e^{j\omega}\right)|$ typically has errors δ_p and δ_s in the passband and stopband (Figure 2.4). Those bands have cutoff frequencies ω_p and ω_s.

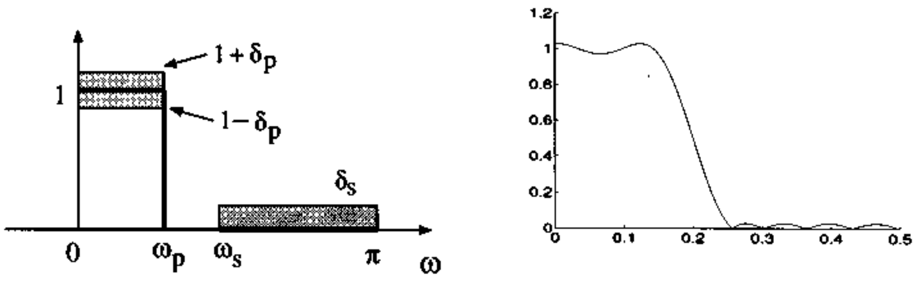


Figure 2.4: Idealized and typical magnitude responses of a lowpass filter.

Given ω_p, ω_s, and N, the errors δ_p and δ_s cannot both be small. Let $W = b/a$ be the relative error weighting. Increasing W in the design algorithm will decrease δ_s and increase δ_p. Figure 2.4 shows a magnitude response plot for a lowpass filter with length 19 and $\omega_p = 0.2\pi$ and $\omega_s = 0.3\pi$. The error weighting W is large (and thus yields a small stopband error).

In the sections below, several methods for the design of FIR digital filters are reviewed. These are based on windowing, minimax criteria, and weighted least squares.

Design by Windowing

The simplest way to truncate the ideal response $\boldsymbol{h}_I$ is by a rectangular window:

$$\boldsymbol{h}(n) = \boldsymbol{h}_I(n)\boldsymbol{w}(n) \quad \text{where} \quad \boldsymbol{w}(n) = \begin{cases} 1; & |n| \le N/2 \\ 0; & \text{otherwise.} \end{cases}$$

This $H\left(e^{j\omega}\right)$ is the best least squares approximation to $H_I\left(e^{j\omega}\right)$. But chopping off the impulse response manifests as passband and stopband ripples in the frequency response. As the window size increases, the ripples get closer to the cutoff frequency ω_c, but these error heights do not decrease. This is the notorious Gibbs phenomenon. (Section 2.2 showed the magnitude response for halfband filters, when $\omega_c = \frac{\pi}{2}$. The solid line has $N = 30$, the dotted line has $N = 10$.) To design FIR filters with better error characteristics, we can smooth out the window $\boldsymbol{w}(n)$:

Hamming window $\quad \boldsymbol{w}(n) = \alpha + (1-\alpha)\cos\left(\frac{2\pi n}{N}\right) \qquad |n| \le \frac{N-1}{2}$

Hanning window $\quad \boldsymbol{w}(n) = \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi n}{N}\right) \qquad |n| \le \frac{N-1}{2}$

Kaiser window $\quad \boldsymbol{w}(n) = \frac{1}{2} I_0\left[\beta\sqrt{1-\left(\frac{2n}{N}\right)^2}\right] / I_0(\beta) \qquad |n| \le \frac{N}{2}$

The parameter β in the Kaiser window controls the attenuation of the lowpass filter. $I_0(x)$ is the modified Bessel function and practical designs use about 20 terms of

$$I_0(x) = 1 + \sum_{k=1}^{\infty}\left[\frac{(0.5x)^2k}{k!}\right]^2.$$

Minimax Criteria (Equiripple Filter)

The filter with the smallest maximum error in passband and stopband is an *equiripple filter*. The equal heights of the ripples (and the number of ripples) assure that the error cannot be reduced—some ripple sizes will go up if others go down. A polynomial of degree N cannot have alternating signs at all ripples. The Remez algorithm to equalize the ripples was adapted to filter design by Parks and McClellan.

Figure 2.5 shows the magnitude response plot of an equiripple lowpass filter. Given a frequency specification in terms of cutoff frequencies (ω_p, ω_s), filter length $(N+1)$ and relative errors (δ_p, δ_s), the equiripple filter has the smallest maximum error in the frequency interval $0 \le \omega \le \pi$. The design algorithm for equiripple filters is the Remez exchange (McClellan-Parks) algorithm.

The order (N) of an equiripple filter is estimated by

$$N \approx \frac{-20\log_{10}\sqrt{\delta_1\delta_2} - 13}{14.6\Delta f} \tag{2.27}$$

where $\Delta f = (\omega_s - \omega_p)/2\pi$ is the transition band.

Equiripple designs are optimal in important respects—*but not optimal for iteration*. The reason is that they have at most one zero at $\omega = \pi$. The sampling operators $(\downarrow 2)$ will mix up the frequency bands that an equiripple filter so carefully separates! *The Daubechies filters go to the other extreme—no ripples at all and maximum flatness at* $\omega = \pi$. Then iteration of

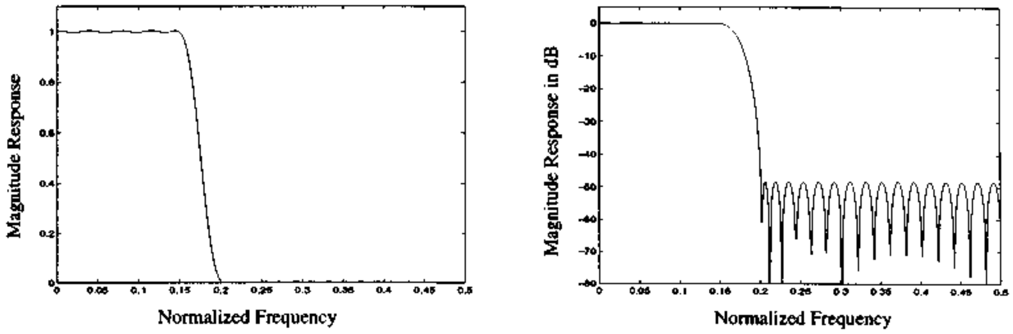


Figure 2.5: Magnitude response of an equiripple filter: normal and dB scales.

$(\downarrow 2)\boldsymbol{H}$ is very stable. But Section 5.5 will show that the transition band $\Delta\omega$ widens from N^{-1} for equiripple to $N^{-\frac{1}{2}}$ for the maxflat Daubechies filters.

A compromise is certainly possible, constraining the minimax design to have a fixed number p of zeros at $\omega = \pi$.

Weighted Least Squares (Eigenfilters)

The eigenfilter approach chooses the filter $\boldsymbol{H}$ to minimize a (weighted) integral error

$$E = \int \left|D(\omega) - H(e^{j\omega})\right|^2 \text{ (weight) } d\omega \qquad (2.28)$$

The integral is over the passband and stopband, not the transition band. $D(\omega)$ is the desired frequency response, possibly one and zero in the two bands. The weighting function is optional. The goal is to express the error as a quadratic form $E = \boldsymbol{h}^T\boldsymbol{P}\boldsymbol{h}$. The unknown filter coefficients are in the vector $\boldsymbol{h}$.

The matrix $\boldsymbol{P}$ is symmetric positive definite because $E > 0$. If the normalization has the form $\boldsymbol{h}^T\boldsymbol{h} = 1$ then the minimization is an eigenvalue problem $\boldsymbol{P}\boldsymbol{h} = \lambda_{\text{min}}\boldsymbol{h}$ in linear algebra:

$$\text{The minimum of } E = \frac{\boldsymbol{h}^T\boldsymbol{P}\boldsymbol{h}}{\boldsymbol{h}^T\boldsymbol{h}} \text{ is } \lambda_{\text{min}}(\boldsymbol{P}).$$

If the normalization is changed to $\boldsymbol{h}^T\boldsymbol{Q}\boldsymbol{h} = 1$, the eigenvalue problem $\boldsymbol{P}\boldsymbol{h} = \lambda\boldsymbol{Q}\boldsymbol{h}$ involves both matrices. When there are several quadratic constraints $\boldsymbol{h}^T\boldsymbol{Q}_k\boldsymbol{h} = 1$, we go beyond an eigenvalue problem — to the Quadratic Constrained Least Squares algorithm. Section 5.4 will apply this QCLS method to filter design. The applications of eigenfilters (one quadratic constraint) are very extensive, and we give two examples: lowpass filters and halfband filters.

Lowpass eigenfilter design: A symmetric filter $\boldsymbol{h}(n) = \boldsymbol{h}(2L-1-n)$ of length $2L$ has response

$$H(e^{j\omega}) = \sum_{n=0}^{2L-1} \boldsymbol{h}(n)e^{-j\omega n} = e^{-j\omega(L-\frac{1}{2})}\sum_{0}^{L-1} \boldsymbol{h}(n)c(\omega).$$

The last sum is $H_{\text{real}}(\omega) = \boldsymbol{h}^T\boldsymbol{c}(\omega)$ with $\boldsymbol{h} = [\boldsymbol{h}(0) \ \cdots \ \boldsymbol{h}(L-1)]$. The vector $\boldsymbol{c}(\omega)$ has components $2\cos(L-\frac{1}{2})\omega, \ldots, 2\cos\frac{\omega}{2}$. In the stopband, from ω_s to π, where $D = 0$ is the desired response, the error is

$$E_{\text{stop}} = \int H_{\text{real}}^2(\omega)d\omega = \boldsymbol{h}^T \int_{\omega_s}^{\pi} \boldsymbol{c}(\omega)\boldsymbol{c}(\omega)^T d\omega\, \boldsymbol{h} \ = \ \boldsymbol{h}^T\boldsymbol{P}_{\text{stop}}\, \boldsymbol{h}.$$

The entries of $\boldsymbol{P}_{\text{stop}}$ are known integrals of cosines. In the passband from 0 to ω_p, we can normalize the desired constant response to be $D = \boldsymbol{h}^T\boldsymbol{c}(0)$. Then the passband error involves the difference between that desired response and the attained response $\boldsymbol{h}^T\boldsymbol{c}(\omega)$:

$$E_{\text{pass}} = \boldsymbol{h}^T \int_0^{\omega_p} (\boldsymbol{c}(0) - \boldsymbol{c}(\omega))(\boldsymbol{c}(0) - \boldsymbol{c}(\omega))^T d\omega\, \boldsymbol{h} \ = \ \boldsymbol{h}^T\boldsymbol{P}_{\text{pass}}\, \boldsymbol{h}.$$

The entries of $\boldsymbol{P}_{\text{pass}}$ are known integrals of $4[1-\cos(n+\frac{1}{2})\omega][1-\cos(m+\frac{1}{2})\omega]$.

We can weight the errors by $E = \alpha E_{\text{stop}} + (1-\alpha)E_{\text{pass}}$. The matrix whose lowest eigenvector is the best $\boldsymbol{h}$ will be $\boldsymbol{P} = \alpha\boldsymbol{P}_{\text{stop}} + (1-\alpha)\boldsymbol{P}_{\text{pass}}$. Figure 2.6 shows the magnitude response of a lowpass filter of length 55 and cutoff frequencies $\omega_p = 0.2\pi$ and $\omega_s = 0.3\pi$. Here, the weight α is 0.5.

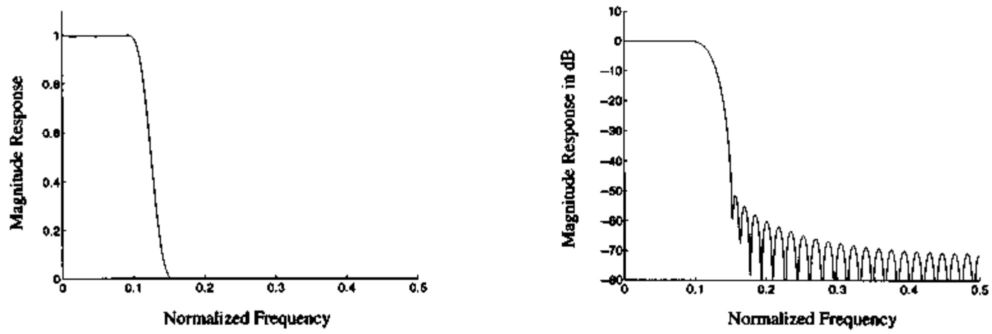


Figure 2.6: Magnitude response plot of a lowpass eigenfilter in normal and dB scale.

Halfband and *M*th-band filter design: In many digital systems, a change in the sampling rate is essential for efficiency in real time. The design of suitable filters for the rate changing operations is important. When the subsampling is by a factor of $M = 2$, we are led to *halfband filters*. When we keep every Mth sample, we need *Mth–band filters*. The center coefficient is $\boldsymbol{h}(0) = 1$ and all coefficients at multiples of M are zero:

$$\textbf{\textit{Mth-band:}} \quad \boldsymbol{h}(nM) = \boldsymbol{\delta}(n) \text{ or in other words } (\downarrow M)\boldsymbol{h} = \boldsymbol{\delta}. \tag{2.29}$$

The filter banks have two channels or M channels. The design begins with a halfband filter or an Mth-band filter. This is the product filter of Section 4.1, which is factored into analysis times synthesis. We shift the filters to make them causal, for implementation. But for design we keep them centered, *and simply zero out the coefficients* $\boldsymbol{h}(nM)$ for $n \neq 0$. This zeros out the corresponding rows and columns of the error matrix $\boldsymbol{P}$ in weighted least squares. The optimum $\boldsymbol{h}$ is the lowest eigenvector of the reduced matrix $\boldsymbol{P}_{\text{red}}$, and this $\boldsymbol{h}$ is Mth–band.

***Design procedure*:** Given N, M, ω_p, ω_s for an Mth-band filter, find $\boldsymbol{P}$ using the eigenfilter formulation. Find $\boldsymbol{P}_{\text{red}}$ by deleting the rows and columns of $\boldsymbol{P}$ that correspond to zero coefficients in $\boldsymbol{h}$. Find the eigenvector $\boldsymbol{h}_{\text{red}}$ corresponding to the minimum eigenvalue of $\boldsymbol{P}_{\text{red}}$. The optimal impulse response $\boldsymbol{h}(n)$ is obtained from $\boldsymbol{h}_{\text{red}}(n)$ by inserting the zero coefficients.

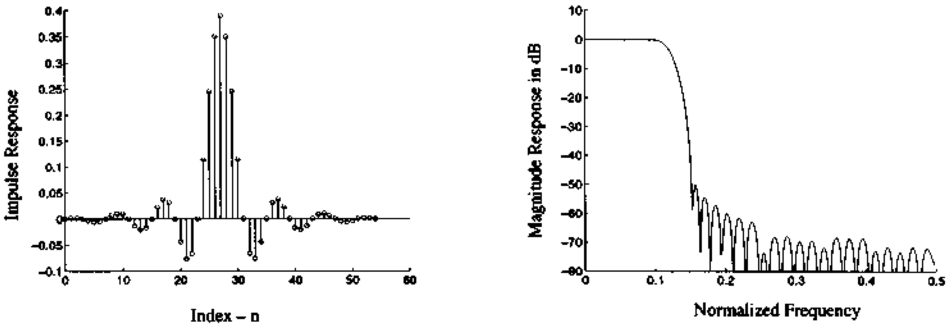


Figure 2.7: Impulse response and magnitude response of a 4th-band FIR filter.

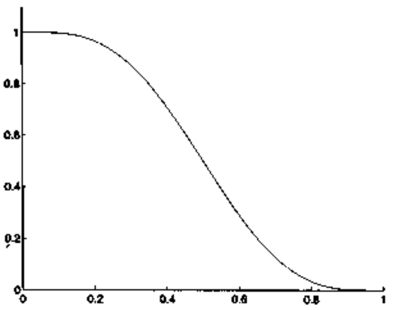
Figure 2.7 shows the magnitude response of an 4th-band filter with $\omega_p = 0.2\pi$ and $\omega_s = 0.3\pi$. Note that $\omega_p + \omega_s = 0.5\pi = \frac{2\pi}{M}$, as required for a symmetric 4th-band filter. The length is 55 (odd length is also required).

Maximally Flat Filter *(Maxflat filter)*: The design of a maxflat filter begins with a maxflat polynomial $\widetilde{H}(y)$. This polynomial of degree $2p - 1$ is determined by p conditions at $y = 0$ and at $y = 1$:

$$\widetilde{H}^{(k)}(0) = \delta(k) \text{ and } \widetilde{H}^{(k)}(1) = 0 \text{ for } 0 \le k < p.$$

The pth order zero at $y = 1$ means that $\widetilde{H}(y) = (1 - y)^p Q(y)$. Then the conditions at $y = 0$ determine $Q(y)$. The details are in Section 5.5, leading to the Daubechies filter by factoring this halfband filter. Here we highlight a remarkable result: $Q(y)$ *consists of the first p terms of the series for* $(1 - y)^{-p}$. With $p = 4$, for example, $\widetilde{H}(y) = (1 - y)^4(1 + 4y + 10y^2 + 20y^3) \approx (1 - y)^4(1 - y)^{-4}$.

Those coefficients 1, 4, 10, 20 come from the binomial series for $(1-y)^{-4}$. They also appear in $(1 + y + y^2 + y^3)^4$. They are binomial numbers. The product $\widetilde{H}(y)$ then has three zero derivatives at $y = 0$. Its graph is in the figure below.



The relations among the variables y and ω and z are

$$\begin{aligned} y &= \frac{1-\cos\omega}{2} &= \tfrac{1}{4}(2-z-z^{-1}) &= -z\Big(\frac{1-z^{-1}}{2}\Big)^2. \\ 1-y &= \frac{1+\cos\omega}{2} &= \tfrac{1}{4}(2+z+z^{-1}) &= z\Big(\frac{1+z^{-1}}{2}\Big)^2. \end{aligned} \qquad (2.30)$$

The zeros at $y = 1$ become zeros at $\omega = \pi$ and at $z = -1$. The flatness at $y = 0$ becomes flatness at $\omega = 0$ and at $z = 1$. We give the frequency response for the centered halfband Daubechies filter:

$$H(e^{j\omega}) = \Big(\frac{1+\cos\omega}{2}\Big)^p \sum_{k=0}^{p-1}\binom{p-1+k}{k}\Big(\frac{1-\cos\omega}{2}\Big)^k. \qquad (2.31)$$

It is a binomial exercise to show that $H = \frac{1}{2}$ at $\omega = \frac{\pi}{2}$. To find $H(z)$, substitute using equations (2.30). Then shift by z^{-p} to make the filter causal.

Note Another form of this polynomial (which is so important in wavelet theory) is the "Bernstein form"

$$H(e^{j\omega}) = \sum_{k=p+1}^{2p+1}\binom{2p+1}{k}\Big(\frac{1+\cos\omega}{2}\Big)^k\Big(\frac{1-\cos\omega}{2}\Big)^{2p+1-k}. \qquad (2.32)$$

As p increases, $H(e^{j\omega})$ approaches the ideal lowpass response (one for $\cos\omega > 0$ and zero for $\cos\omega < 0$). The $p^{-1/2}$ width of the transition band is established in Section 5.5, together with the approximation of the zeros of $H(z)$.

The ideal is infinitely flat. Of course it needs infinitely many coefficients.

Problem Set 2.3

1. Truncate the ideal lowpass filter after three nonzero coefficients. What is this windowed filter $h(n)$? Sketch the graph of $H(\omega)$.
2. Truncate the ideal lowpass filter after 4 terms and 8 terms. Draw the frequency responses $H(\omega)$ to see the Gibbs phenomenon.
3. Compare the graphs of the ideal brick wall filter truncated after 20 terms (rectangular window) and the Kaiser window $w(n)$. Choose a suitable Kaiser parameter β. What are the maximum errors in the passband?
4. Use MATLAB to design equiripple halfband filters of length 8 and 20. Compute the height of the ripples.
5. What is the frequency response for the maxflat Daubechies filter (2.31) with $p = 2$? Graph $H(\omega)$ by hand or by computer. What are its symmetries?
6. Graph the maxflat Daubechies filter response (2.31) for $p = 8$. What are the differences from the truncated ideal filter and the equiripple filter?
7. Construct a 4-tap lowpass filter that you approve of. What properties have you achieved?
8. Derive the Bernstein form (2.32) of the Daubechies polynomial $H(e^{j\omega})$ from her original form (2.31).
9. Formulate the eigenfilter design for a highpass filter with cutoff frequency ω_c.
10. Compute $h(n)$ for the halfband Daubechies filter with $p = 5$. Verify that $H(e^{j\omega})$ has four zero derivatives at $\omega = 0$ and $\omega = \pi$.

2.4 Fourier Analysis

This section might be called "Notes on Fourier Analysis." The subject is enormous — too large for our book! We pick out key points that are needed for signal processing and for wavelets.

The Fourier transform of a signal $x(n)$ is a function $X(\omega)$. The signal is in the time domain, $X(\omega)$ is in the frequency domain. The time variable n is discrete, the frequency variable ω is continuous. $X(\omega)$ is 2π-periodic because each exponential $e^{in\omega}$ is 2π-periodic:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-in\omega}. \tag{2.33}$$

Fourier analysis studies the connections between $x(n)$ and $X(\omega)$ — how the properties of the signal are reflected in its transform. The ***inverse Fourier transform*** recovers $x(n)$:

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)e^{in\omega}d\omega. \tag{2.34}$$

This formula synthesizes x by combining the complex exponentials. To find $x(N)$, *multiply equation* (2.33) *by* $e^{iN\omega}$ *and integrate from* $-\pi$ *to* π. The integral of $e^{-in\omega}$ times $e^{iN\omega}$ is zero except when $n = N$. That integral is 2π, leading to (2.34).

Fourier analysis usually starts with $f(t)$ and computes its coefficients. Signal processing starts with the coefficients $x(n)$ and transforms to $X(\omega)$.

Note about orthogonality. Real vectors are orthogonal (perpendicular) when $x \cdot y = 0$. Real functions are orthogonal when $\int X(\omega)Y(\omega)d\omega = 0$. If the vectors or the functions are complex, there is a small but important change. We take complex conjugates of one vector (say x, in the physics convention) and of one function $X(\omega)$. *Orthogonality in the complex case means*

$$\bar{x} \cdot y = \sum \overline{x(n)}y(n) = 0 \quad \text{and} \quad (X, Y) = \int_{-\pi}^{\pi} \overline{X(\omega)}Y(\omega)\,d\omega = 0. \tag{2.35}$$

The integration $\int e^{-in\omega}e^{iN\omega}\,d\omega = 0$ does not say that $e^{-in\omega}$ is orthogonal to $e^{iN\omega}$. It says that $e^{in\omega}$ is orthogonal to $e^{iN\omega}$ (for $N \neq n$). It is this orthogonality that allows the inverse transform to have the clean and simple formula (2.34).

The discrete analogue of an orthonormal transform is a square matrix with orthonormal columns. This is an "orthogonal" matrix if real, a "unitary" matrix if complex. For such a matrix U, the inverse is again clean and simple. It equals the conjugate transpose $\overline{U}^T$. This applies to orthonormal filter banks, when the rows of $(\downarrow 2)\boldsymbol{H}_0$ and $(\downarrow 2)\boldsymbol{H}_1$ are orthonormal. Their transposes are the columns of $\boldsymbol{F}_0(\uparrow 2)$ and $\boldsymbol{F}_1(\uparrow 2)$ — also orthonormal.

Is there a connection between discrete and continuous orthogonality? If two signals are orthogonal (in discrete time), are their transforms orthogonal (in continuous frequency)? The question is important and the answer is ***yes***.

This answer follows from the orthogonality of the exponentials $e^{in\omega}$ — on which the whole theory depends. For any x and y, the inner products in the time and frequency domains are equal up to a factor of 2π:

$$2\pi \sum_{n=-\infty}^{\infty} \overline{x(n)}y(n) = \int_{-\pi}^{\pi} \overline{X(\omega)}Y(\omega)\,d\omega. \tag{2.36}$$

For proof, substitute $\sum \overline{x(N)}e^{iN\omega}$ for $\overline{X(\omega)}$ on the right. Also substitute $\sum y(n)e^{-in\omega}$ for $Y(\omega)$. Integrate from $-\pi$ to π. By orthogonality, the only terms with nonzero integrals are those with $N = n$. The integral is 2π, multiplied by the number $\overline{x(n)}y(n)$. The left side of (2.36) is the sum of these nonzero terms.

Special case when $x = y$. Now the two vectors are the same. Their transforms are the same. We are integrating $\overline{X(\omega)}X(\omega)$, which is $|X(\omega)|^2$, because $a + ib$ times its conjugate $a - ib$ is $a^2 + b^2$. For the same reason $\overline{x(n)}x(n) = |x(n)|^2$. With $x = y$, the energy in the signal (times 2π) equals the energy in the transform:

$$2\pi \sum_{-\infty}^{\infty} |x(n)|^2 = \int_{-\pi}^{\pi} |X(\omega)|^2 \, d\omega. \tag{2.37}$$

Example 2.8. Suppose x is $(1, \beta, \beta^2, \ldots)$. Its energy is $1 + \beta^2 + \beta^4 + \cdots = 1/(1 - \beta^2)$. Its transform is a one-sided sum, in this case a geometric series:

$$X(\omega) = \sum_{0}^{\infty} \beta^n e^{-in\omega} = 1 + \beta e^{-i\omega} + \left(\beta e^{-i\omega}\right)^2 + \cdots = \frac{1}{1 - \beta e^{-i\omega}}.$$

Consider the inverse transform from $X(\omega)$ back to $x(n)$:

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) e^{in\omega} d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{in\omega} d\omega}{1 - \beta e^{-i\omega}}.$$

How do we see that this integral gives the correct signal $(1, \beta, \beta^2, \ldots)$? The direct way is to write the integrand as $e^{in\omega}\left(1 + \beta e^{-i\omega} + \beta^2 e^{-2i\omega} + \cdots\right)$. Integration picks out the correct power β^n. (If $n < 0$ then the integral gives zero.) The indirect way is to substitute z for $e^{i\omega}$, and integrate $z^n/(z - \beta)$ around the unit circle:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{in\omega} d\omega}{1 - \beta e^{-i\omega}} = \frac{1}{2\pi i} \int_{|z|=1} \frac{z^n (dz/z)}{1 - (\beta/z)}.$$

There is a pole at $z = \beta$. The *residue* at this pole is β^n. This is the answer we want. Again the case $n < 0$ is separate (with two poles) and gives zero.

The actual calculation of such an inversion integral is generally difficult or impossible. We seldom need to do it. For a ratio of polynomials it can be done in an emergency by the residue method of complex integration.

Example 2.9. An allpass filter has $|H(\omega)| = 1$ so that $|Y(\omega)| = |X(\omega)|$. The filter conserves energy. The integral of $|Y(\omega)|^2$ equals the integral of $|X(\omega)|^2$, because these functions are the same:

$$\textit{Allpass:} \quad \int_{-\pi}^{\pi} |Y(\omega)|^2 d\omega = \int_{-\pi}^{\pi} |H(\omega)X(\omega)|^2 d\omega = \int_{-\pi}^{\pi} |X(\omega)|^2 d\omega.$$

Therefore ***the output energy equals the input energy:***

$$\sum_{-\infty}^{\infty} |y(n)|^2 = \sum_{-\infty}^{\infty} |x(n)|^2 \quad \text{or} \quad \|y\|^2 = \|x\|^2. \tag{2.38}$$

Please do not think that each $|y(n)| = |x(n)|$. It is the *frequency* response that has $|H(\omega)| = 1$. The energy in each frequency band is conserved by an allpass filter, not the energy in each time band.

Convergence of the Fourier Series

In defining $X(\omega)$, we have been assuming that this series converges. Otherwise what meaning do we give to $X(\omega)$? We are touching here on a central problem of mathematical analysis (with a literature that goes back for centuries). Touching this problem is as much as we can do — by identifying three types of convergence, and the signals that produce each type.

As with all infinite series, a few terms have nothing to do with convergence. By changing a finite number of inputs $x(n)$ we certainly change the transform — but we do not alter convergence or divergence. It is the behavior of $x(n)$ *for large n* that is crucial. Here are three types of convergence:

1. Uniform convergence with $\sum |x(n)| < \infty$
2. Strong convergence (in L^2) with $\sum |x(n)|^2 < \infty$
3. Weak convergence allowing polynomial growth in $x(n)$.

1. Uniform convergence Suppose the magnitudes $|x(n)|$ have a finite sum. These are also the magnitudes of $x(n)e^{-in\omega}$, because $|e^{-in\omega}| = 1$. Those terms have different phases, which may produce cancellation when we add them. When $\sum |x(n)|$ converges, *we don't need that help.* The series of magnitudes converges "absolutely," and the series $\sum x(n)e^{-in\omega}$ converges "uniformly." Then the sum $X(\omega)$ is a *continuous function* — with no jumps.

In the example with $x = (1, \beta, \beta^2, \ldots)$ we imposed $|\beta| < 1$. That produced uniform convergence. The transform $X(\omega) = 1/\left(1 - \beta e^{-i\omega}\right)$ is a continuous function. But not all continuous functions have $\sum |x(n)| < \infty$.

In the brick wall filter, the odd coefficients have magnitude $|h(n)| = \frac{1}{\pi n}$. The sum of magnitudes does *not* converge. The terms $\frac{1}{\pi n}$ do not go to zero quickly enough. The sum $H(\omega) = \sum h(n)e^{-in\omega}$ does not converge uniformly. And, in fact, $H(\omega)$ is a step function (or square wave) with a jump.

2. Convergence in energy (L^2 convergence) Suppose the *squared* magnitudes $|x(n)|^2$ have a finite sum. By squaring, small terms become much smaller. Convergence is easier to achieve. If the sum of $|x(n)|$ is finite, the sum of squares is certainly finite. Then the Fourier series converges in L^2. This "squared" test is passed by the Fourier series for a step function:

$$|h(n)| = \frac{1}{\pi n} \quad \text{has} \quad \sum |h(n)|^2 = \sum \frac{1}{\pi^2 n^2} = \text{convergent series.}$$

Comparing the sums of $1/n$ and $1/n^2$ is like comparing the integrals of $1/x$ and $1/x^2$. One integral is $\log x$, which becomes large as $x \to \infty$. The area under $1/x^2$ stays finite as $x \to \infty$.

When the squared magnitudes $|x(n)|^2$ have a finite sum, the squared magnitude $|X(\omega)|^2$ has a finite integral. They are equal apart from 2π. Then $X(\omega)$ is a function in the Hilbert space denoted by L^2, just as $x(n)$ is a vector in the Hilbert space denoted by ℓ^2. These spaces contain all functions and vectors with finite energy: *the square of the L^2 norm or the ℓ^2 norm is the energy.*

Functions in Hilbert space may have jumps. Those jumps have no energy (no contribution to the integral). The *derivative* of a jump is a Dirac delta function. Its coefficients are all $x(n) = \frac{1}{2\pi}$ and its energy is infinite, so this function is outside the space L^2.

But the delta function is included in the third type of convergence.

3. Weak convergence (to a distribution) Distributions $F(\omega)$ are defined by their inner products with smooth functions $G(\omega)$. *For $\int F(\omega)G(\omega)\,d\omega$, use integration by parts.* We integrate $F(\omega)$, which needs it, and we differentiate $G(\omega)$, which can take it. The indefinite integral of the delta function $F(\omega) = \delta(\omega)$ is the unit step $H(\omega)$. The definite integral of $\delta(\omega)G(\omega)$ is the number $G(0)$:

$$\int_{-\pi}^{\pi} \delta(\omega)G(\omega)\,d\omega = [H(\omega)G(\omega)]_{-\pi}^{\pi} - \int_{-\pi}^{\pi} H(\omega)G'(\omega)\,d\omega = G(\pi) - \int_{0}^{\pi} G'(\omega)\,d\omega = G(0).$$

This defines $\delta(\omega)$. It is a distribution, a derivative of a true function. Still it has Fourier coefficients that are easy to find, with $G(\omega) = e^{in\omega}$:

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \delta(\omega)e^{in\omega}\,d\omega = \frac{1}{2\pi}\cdot 1.$$

All frequencies are present in the same amount. The Fourier series is

$$\delta(\omega) = \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} e^{in\omega}. \tag{2.39}$$

On the left is a distribution. On the right is a divergent series. The terms don't even approach zero. But in a weak sense those terms cancel each other to produce zero, away from the spike at $\omega = 0$ where they reinforce.

Weak convergence is based on the same idea of testing inner products with smooth $G(\omega)$. The series *converges weakly*, say at $\omega = 0$:

$$G(0) = \int_{-\pi}^{\pi} \delta(\omega)G(\omega)\,d\omega = \frac{1}{2\pi}\sum_{n}\Big(\int_{-\pi}^{\pi} e^{in\omega}G(\omega)\,d\omega\Big)e^{in0}. \tag{2.40}$$

You see the Fourier coefficients of $G(\omega)$ on the right side, adding to $G(0)$.

Numerically, this weak convergence is not so great. Figure 2.8 shows the sum of 41 terms. In a pointwise sense, and in area, those side lobes will not shrink to zero! In the L_2 sense, the energy is growing with every term. But in a weak sense, this sum is approaching $\delta(\omega)$. As the number of terms increases, the oscillations become faster (not smaller). Multiplied by a smooth $G(\omega)$, the main central lobe picks out $G(0)$ and the integral over the oscillations approaches zero.

Question: What is the weak limit of pure oscillations $e^{in\omega}$ as $n \to \infty$?

Answer: The limit is the zero function. The inner products $\int e^{in\omega}G(\omega)\,d\omega$ approach zero. Oscillations converge *weakly* to their average value.

Question: Does the series $\delta'(\omega) = \frac{i}{2\pi}\sum n\, e^{in\omega}$ for the derivative of a Dirac function also converge weakly? The Fourier coefficients are growing with n.

Answer: Yes. Integrate again by parts. The integral of $\delta'(\omega)G(\omega)$ is $-G'(0)$.

The ***Gibbs phenomenon*** is just the integral of $\sum e^{in\omega}$. The integral of $\delta(\omega)$ is a step function. The integral of $e^{in\omega}$ introduces $\frac{1}{n}$, so there is L^2 convergence to the step (but not uniform convergence). With integration, *the area under the side lobes become crucial.* The area shows as a height in the Gibbs figure (the integral of the delta function is a step). This undesirable oscillation of Fourier series at a jump (an *edge* in image processing) has brought forward localized bases like wavelets.

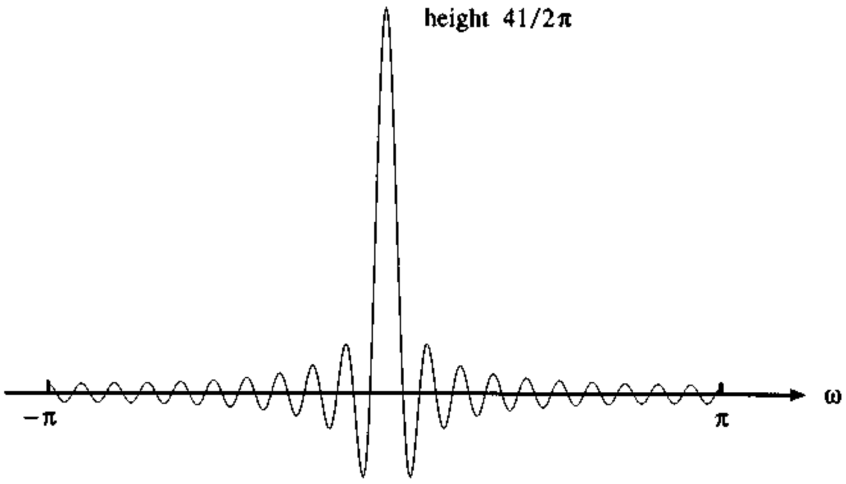


Figure 2.8: The Fourier series $\frac{1}{2\pi}\sum e^{-in\omega}$ converges weakly to $\delta(\omega)$. Its integral shows the Gibbs phenomenon (nonuniform convergence).

It is significant that the same Gibbs difficulty in shock calculations has been handled very differently. The finite difference schemes are made non-oscillatory by *nonlinear terms*. The nonlinearity is active where it is needed. It is inactive where the solution is smooth. Morel's book [Mo] gives a strong impetus to this nonlinear idea for image processing.

For iteration of the filter $\boldsymbol{h} = (\frac{1}{2}, 0, 0, \frac{1}{2})$, we will see weak convergence in Chapter 7. The cascade algorithm (= lowpass iteration) produces functions that oscillate between 1 and 0 on the interval $[0, 3]$. They have no limit in L^2. The weak limit of the oscillations is a stretched box $\phi(t)$ with constant value $\frac{1}{3}$.

Good filters give L^2 convergence (and usually uniform convergence) when iterated. The necessary and sufficient Condition E is in Section 7.2. But a "good filter" in the classical sense, with small errors in the passband and stopband, may fail in iteration!

The requirement for success in iteration is flatness of the response $H(\omega)$. *The number of zeros at $\omega = \pi$ is absolutely critical.* This is the new property that has become important. It makes the iteration process strong and not weak, regular and not oscillatory. It must be built in or iterations will oscillate — as happened with highly regarded filters.

Poisson's Summation Formula

The Fourier coefficients of the Dirac delta function on $[-\pi, \pi]$ are all $\frac{1}{2\pi}$. By periodicity, the Dirac delta becomes a *Dirac comb*. There is an impulse at every multiple of 2π. The Fourier series for this periodic train of delta functions is

$$\sum_{k=-\infty}^{\infty} \delta(\omega - 2\pi k) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\omega}. \tag{2.41}$$

The left side is a sum of impulses. The right side is its Fourier series, *converging weakly* as above. The formula is often seen with t instead of ω, because Fourier analysis usually starts with continuous time. Then the frequencies are discrete. In signal processing it is the other way around.

Equation (2.41) is a short statement of the Poisson formula, but it involves weak convergence. The terms don't approach zero. To see ordinary convergence we take inner products with any smooth function $G(\omega)$. With integrals over the whole line, we ask $G(\omega)$ to decay as $|\omega| \to \infty$. The inner product with delta functions gives point values $G(2\pi k)$. The inner product on the right gives point values $\widehat{G}(n)$ of the Fourier transform. The formula says that the two sums are equal:

$$\textbf{\textit{Poisson's Summation Formula}} \qquad \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} G(2\pi k) = \sum_{n=-\infty}^{\infty} \widehat{G}(n). \tag{2.42}$$

This is a remarkable formula. It relates the samples of G to the samples of its transform $\widehat{G}$. We can change the spacing of one set of samples to any $T > 0$, provided the other set of samples is spaced at $\frac{2\pi}{T}$.

Smoothness of $X(\omega)$ and Decay of $x(n)$

The regularity of a function is partly revealed in its Fourier coefficients. When $X(\omega)$ is continuous, its Fourier coefficients $x(n)$ approach zero. But the coefficients can approach zero when $X(\omega)$ is *not* continuous; the step function is an example. There is a gap between $x(n) \to 0$ and $\sum |x(n)| < \infty$, which is at the heart of Fourier analysis:

$$\sum |x(n)| < \infty \qquad \Longrightarrow \qquad \text{continuous } X(\omega) \qquad \Longrightarrow \qquad x(n) \to 0.$$

One virtue of wavelets is that such gaps can be closed. Instead of necessary conditions for smoothness, and then sufficient conditions for smoothness, we can find *necessary and sufficient conditions*. These are conditions on the wavelet coefficients $x(n)$, for the function $X(\omega)$ to have specified regularity. $X(\omega)$ lies in a specified function space when $x(n)$ lies in a corresponding vector space [DeLu]. The function norm and vector norm are equivalent — as for $X(\omega)$ in L^2 and $x(n)$ in ℓ^2 by Parseval's formula.

Fourier coefficients give partial information from the magnitudes $|x(n)|$. Each extra order of smoothness in $X(\omega)$ is reflected in one extra order of decay in $|x(n)|$.

Theorem 2.2 *Suppose $X(\omega)$ has s continuous derivatives. Then $n^s|x(n)| \to 0$ as $|n| \to \infty$.*

When $s = 0$ this is the Riemann-Lebesgue Lemma. The Fourier coefficients approach zero (we really only need $\int |X(\omega)|d\omega$ to be finite). For integers $s = 1, 2, 3, \ldots$ we look at the Fourier coefficients of the derivative $X^{(s)}(\omega)$. Those coefficients are $(in)^s x(n)$. They approach zero (again by Riemann-Lebesgue) when $X^{(s)}(\omega)$ is continuous. This is the theorem.

In compression of a signal, this decay of coefficients is crucial. Small coefficients are removed (partly or completely). For a Fourier basis, the smoothness of an image determines the decay of coefficients. For a wavelet basis, it is the *piecewise smoothness* that matters. Wavelets are well adapted to piecewise smooth functions (with edges). Wavelets are local where Fourier waves are global.

The theorem was stated for integers $s = 0, 1, 2, \ldots$ but fractions can be allowed. This is important for a satisfactory theory, because functions like $X(\omega) = |\omega|^{1/2}$ are more than continuous ($s = 0$) and less than differentiable ($s = 1$). The in-between smoothness is measured by the *Hölder exponent*. The function $X(\omega)$ belongs to the space C^s, $0 < s < 1$, if it satisfies the Hölder condition

$$|X(\omega_2) - X(\omega_1)| \leq \text{ constant } |\omega_2 - \omega_1|^s .$$

The square root function $|\omega|^{1/2}$ belongs to $C^{1/2}$. Similarly $|\omega|^s$ belongs to C^s. For $s > 1$ we would take $[s]$ derivatives first. The Hölder exponent of that derivative is the fractional part $s - [s]$. Then the Fourier coefficients decay to order s at least: $n^s|x(n)| \to 0$.

The Fourier basis automatically takes advantage of high regularity. *The wavelet basis takes full advantage of a high s, only if it is designed to do so.* The lowpass filter must have more than s zeros at π. Haar wavelet expansion does not converge faster as the function gets smoother. Therefore Haar compression is poor.

For biorthogonal wavelets, the analyzing filter $\boldsymbol{H}_0$ governs the decay of coefficients. The synthesizing filter $\boldsymbol{F}_0$ governs the smoothness of the output. *We would like both to have many zeros at π!* When forced to choose, whether to put zeros into $\boldsymbol{H}_0$ or $\boldsymbol{F}_0$, the preference often goes to $\boldsymbol{F}_0$ — then the synthesis basis functions will be smooth.

We emphasize one more point. *In the L^2 norm,* where "mean square" replaces "maximum," *the Fourier coefficients exactly reflect the smoothness.* The coefficients are in ℓ^2 precisely when the function is in L^2. The energy is the same for both, by Parseval's identity. This equality extends to the sth derivative $X^{(s)}(\omega)$, whether s is an integer or not:

$$\int_{-\pi}^{\pi} \left|X^{(s)}(\omega)\right|^2 d\omega = 2\pi \sum_{-\infty}^{\infty} \left|n^s x(n)\right|^2. \tag{2.43}$$

When s is not an integer, we *define* this derivative $X^{(s)}(\omega)$ by its coefficients $(in)^s x(n)$. The equation above is just Parseval itself, for the derivative. Wavelet theory is greatly simplified by working in the Hilbert spaces of functions with s derivatives in L^2, rather than the Hölder spaces C^s of functions with s continuous derivatives.

The number p of zeros at π is crucial in both cases. For L^2 spaces we will find in Chapter 7 the exact smoothness of $\phi(t)$ and $w(t)$.

Heisenberg's Uncertainty Principle

The underlying property of wavelets is that they are pretty well localized in both time and frequency. The functions $e^{i\omega t}$ are perfectly localized at ω but they extend over all time. Wavelets are *not* at a single frequency, or even a finite range, but they are limited to finite time. As we rescale, the frequency goes up by 2^j and the time interval goes down by 2^j. This suggests that the *product* of frequency interval and time interval is a stable quantity. The Heisenberg Uncertainty Principle makes those definitions precise, and gives a *lower bound* for the product.

We emphasize that the wavelet can only be "pretty well localized." It cannot have finite support in both t and ω. (A famous theorem.) Instead of the support length we use the variances

$$\sigma^2 = \int_{-\infty}^{\infty} t^2|f(t)|^2\,dt \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega^2|\widehat{f}(\omega)|^2\,d\omega. \tag{2.44}$$

We are no longer working with periodic functions, so t and ω extend from $-\infty$ to ∞.

Theorem 2.3 (Heisenberg) *If $\|f\| = 1$ then the product $\sigma\widehat{\sigma}$ is at least $\frac{1}{2}$.*

The lower bound $\frac{1}{2}$ is attained by the Gaussian function $f(t) = e^{-t^2}$. Its transform $\widehat{f}(\omega)$ is also a Gaussian. These have infinite support! But they are as local as possible, measured by σ and $\widehat{\sigma}$. We can rescale time by c and frequency by $\frac{1}{c}$, we can shift to $t - t_0$, and we can modulate by $e^{i\omega_0 t}$. The variances σ^2 and $\widehat{\sigma}^2$ would be computed around t_0 and ω_0, and the bound $\frac{1}{2}$ is still reached.

This led Gabor to use Gaussians in constructing "time-frequency atoms." But numerically, finite support in time is better.

Proof of the Heisenberg Principle. The key is that $\sigma = \|tf(t)\|$ and $\widehat{\sigma} = \|f'(t)\|$. The principle applies to pairs of operators, like position $\boldsymbol{P}$ and momentum $\boldsymbol{Q}$, that have the property $\boldsymbol{QP}-\boldsymbol{PQ} = \boldsymbol{I}$. When this property holds, the proof comes directly from the Schwarz inequality:

$$1 = \|f\|^2 = (f, (\boldsymbol{QP} - \boldsymbol{PQ})f) \le 2\ \|\boldsymbol{Q}f\|\ \|\boldsymbol{P}f\| = 2\widehat{\sigma}\sigma. \tag{2.45}$$

In our case $\boldsymbol{P}$ is multiplication by t and $\boldsymbol{Q}$ is differentiation:

$$(\boldsymbol{QP} - \boldsymbol{PQ})f(t) = \frac{d}{dt}\big(tf(t)\big) - t\frac{df}{dt} = \boldsymbol{I}f(t) \quad \text{as required.}$$

Problem 4 develops this proof directly from the definitions of σ and $\widehat{\sigma}$.

Problem Set 2.4

1. Add the terms $\displaystyle\frac{1}{2\pi}\sum_{-N}^{N} e^{in\omega}$. This is a partial sum of the Fourier series for $\delta(\omega)$. At what ω does it come down to zero, at the end of the main lobe in Figure 2.8? Where is the end of the first side lobe?
2. Describe a continuous function $X(\omega)$ whose coefficients have infinite sum $\sum |x(n)| = \infty$.
3. What is Poisson's summation formula for the Gaussian $G(\omega) = e^{-\omega^2}$?
4. Heisenberg's Uncertainty Principle is $\sigma\widehat{\sigma} \ge \frac{1}{2}$. The Schwarz inequality gives
$$\left|\int tf(t)f'(t)\,dt\right|^2 \le \int |tf(t)|^2\,dt \int |f'(t)|^2\,dt.$$
Identify the right side as $\sigma^2\widehat{\sigma}^2$. Integrate by parts on the left side to get
$$\int_{-\infty}^{\infty} t\left[f(t)f'(t)\right]dt = t\frac{f(t)^2}{2}\bigg]_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\frac{f(t)^2}{2}\,dt.$$
If the integrated part is zero at $\pm\infty$, and if $\|f\| = 1$, deduce that $\sigma\widehat{\sigma} \ge \frac{1}{2}$. Equality holds when $tf(t)$ is proportional to $f'(t)$, which leads to Gaussians $f(t) = e^{-ct^2}$.
5. What power of $\frac{1}{n}$ gives the decay rate for the coefficients $x(n)$ of
    1. $X(\omega) = \omega =$ discontinuous function at $\omega = \pm\pi$.
    2. $X(\omega) = |\omega| =$ continuous function with period 2π.
    3. $X(\omega) = \omega^2 =$ continuous function with period 2π.
    4. $X(\omega) =$ spline with jump in the third derivative $X'''(\omega)$.
6. Compute the Fourier coefficients and the energy for

    (a) $X(\omega) = \dfrac{1}{1 - \frac{1}{2}e^{-i\omega}}$ (b) $X'(\omega)$

    (c) Periodic box function $H(\omega) = \begin{cases} 1 & 0 \le \omega < \pi \\ 0 & -\pi \le \omega < 0 \end{cases}$
7. Show that $X(\omega) = (1-\beta e^{-i\omega})^{-1}$ has the same energy as $\boldsymbol{x} = (1, \beta, \beta^2, \ldots)$. Use a substitution in the integral of $|X(\omega)|^2$.
8. Show that the Heisenberg product $\sigma\widehat{\sigma}$ is not changed by *dilation*, *modulation*, and *translation*: $f(t)$ is transformed to $2^{j/2}f(2^j t)$ and $e^{i\omega t}f(t)$ and $f(t - s)$.
9. Determine the Fourier transform of the signal $x(n) = \alpha^{|n|}$, $|\alpha| < 1$.

2.5 Bases and Frames

Above all, this book is about **the choice of a good basis.** In reality, that choice governs everything. Every transform is a change to a different basis. The contribution of filter banks and wavelets (and Fourier transforms and local cosines and wavelet packets) is *to offer new bases.* The whole subject has been reopened, by the demands of its applications.

What is a basis and what makes it good? A basis is a sequence of vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots$ or functions $v_1(t), v_2(t), \ldots$ with the property of *unique representation*:

> Every vector $\boldsymbol{v}$ or function $v(t)$ in the space can be represented
> in one and only one way as $\boldsymbol{v} = \sum b_i \boldsymbol{v}_i$ or $v(t) = \sum b_i v_i(t)$.

There is exactly one representation of every vector and function. The zero vector and zero function can only be represented with all $b_i = 0$. The basis functions are *linearly independent.*

It is usual to require convergence in norm, $\|\boldsymbol{v} - \sum_1^N b_i \boldsymbol{v}_i\| \to 0$ as $N \to \infty$. For L^2 spaces this norm is the square root of the energy. Section 1.5 mentions four reasons for that choice.

There are two separate properties to be established for any proposed basis: (1) *linear independence* and (2) *completeness.* Adding extra vectors will destroy independence. Removing vectors from the basis will kill completeness. *Linear independence is automatic for orthonormal vectors.* When all angles are 90° and all lengths are unity, there is no chance of degeneracy. In infinite dimensions we meet the possibility that angles can approach zero without reaching zero. In that case the basis is unstable. The coefficients b_i are out of control. In the good case, the coefficients satisfy

$$A\,\|\boldsymbol{v}\|^2 \le \sum |b_i|^2 \le B\,\|\boldsymbol{v}\|^2 \quad \text{with} \quad A > 0. \tag{2.46}$$

This is the defining property of a *Riesz basis,* also known as a *stable basis* or an *unconditional basis.* A key assumption in wavelet theory is that the translates $\phi(t-n)$ of the scaling function are a Riesz basis (for the space V_0 inside L^2). We will find the test to be applied to $\widehat{\phi}(\omega)$, and the equivalent Condition E on the filter coefficients, to produce a Riesz basis of scaling functions.

Dual Bases and Dual Frames

In a few lines, we can give the main points about bases and frames. These ideas will be developed below — bases will come first. Here are the key points in a hurry:

Dual bases come from columns $\boldsymbol{v}_n$ of $\boldsymbol{T}^{-1}$ and rows $\boldsymbol{r}_n$ of $\boldsymbol{T}$: $\boldsymbol{T}^{-1}\boldsymbol{T} = \boldsymbol{I}$

Dual frames come from columns $\boldsymbol{v}_n^+$ of $\boldsymbol{T}^+$ and rows $\boldsymbol{r}_n$ of $\boldsymbol{T}$: $\boldsymbol{T}^+\boldsymbol{T} = \boldsymbol{I}$.

The difference is that *the frame vectors $\boldsymbol{v}_n^+$ need not be independent.* They can be redundant (Figure 2.9). We still require the coefficients $b_i = \langle \boldsymbol{r}_i, \boldsymbol{v}\rangle$ to satisfy (2.46), but other combinations of the $\boldsymbol{v}_n^+$ can reconstruct the same $\boldsymbol{v}$.

$\boldsymbol{T}^+$ is only a left-inverse for a frame. It becomes a two-sided inverse for a basis.

In N-dimensional space, frames contain $M > N$ vectors. The matrix $\boldsymbol{T}$ is $M \times N$, and its left-inverse $\boldsymbol{T}^+$ is $N \times M$. ***The equation $\boldsymbol{T}\boldsymbol{T}^+ = \boldsymbol{I}$ is not true.*** For finite dimensions the theory and examples are particularly clear — our eventual applications are to infinite dimensions.

Note 1 *Change of Basis* Suppose $\boldsymbol{T}$ is a bounded linear operator with a bounded inverse. If the sequence $\{\boldsymbol{v}_n\}$ is a Riesz basis, so is the sequence $\{\boldsymbol{T}^{-1}\boldsymbol{v}_n\}$. When we expand $\boldsymbol{T}\boldsymbol{v} = \sum c_n \boldsymbol{v}_n$ in the

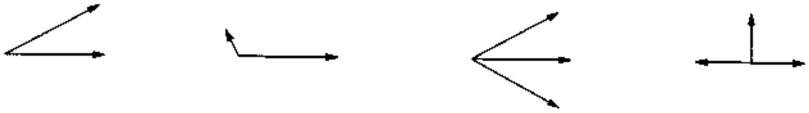


Figure 2.9: Good basis, bad basis, good frame, bad frame, all in $\mathbf{R}^2$.

original basis and multiply by T^{-1}, this gives the expansion $v = \sum c_n(T^{-1}v_n)$ in the new basis: $A = B = 1$. The new Riesz constants A, B come from the original A, B and the norms of T and T^{-1}.

The classical transforms of mathematics have a stronger property than invertibility. *T is often a unitary operator.* The inverse of T becomes the conjugate transpose T^*. The norms are $\|T\| = \|T^{-1}\| = 1$. If the original basis is orthonormal so is the new basis: $A = B = 1$. The Fourier transform and wavelet transform are unitary operators — when we use orthonormal wavelets! Biorthogonal wavelets lead to non-unitary operators T and to Riesz bases — but not orthonormal. Fortunately $\|T\|$ and $\|T^{-1}\|$ are in practice surprisingly close to 1.

In finite dimensions, every basis is a Riesz basis. *The basis vectors for $\mathbf{R}^N$ are the columns of an invertible $N \times N$ matrix.* That matrix is T^{-1}. It is essential to distinguish the change of coordinates (which uses T) from the new basis (the columns of T^{-1}). The fundamental fact is $T^{-1}T = I$. We use it now, multiplying *columns times rows*:

$$v = T^{-1}Tv = \sum_{n=1}^{N}(\text{column } n \text{ of } T^{-1})(\text{row } n \text{ of } T)v$$

The nth basis vector is $v_n = \text{column } n \text{ of } T^{-1}$

The nth coordinate is $b_n = (\text{row } n \text{ of } T)v = \langle r_n, v\rangle$.

Those three lines give $v = \sum b_n v_n$.

Every $N \times N$ matrix is a bounded operator. When the inverse exists, it is also bounded. But infinite matrices can represent unbounded operators. The averaging filter with coefficients $\frac{1}{2}, \frac{1}{2}$ is bounded. The inverse filter with coefficients $2, -2, 2, -2, 2, -2, \ldots$ is unbounded. The Riesz requirement, that T and T^{-1} are both bounded, is needed for a stable change of basis. An orthonormal basis ($A = B = 1$) has condition number 1. Then the product $\|T\|\,\|T^{-1}\|$ is the **condition number** of the new basis.

Bases from Filter Banks

The important examples for us are the bases from filter banks (discrete time) and the bases from wavelets (continuous time). Chapter 1 gave an example of both. The discrete time basis vectors had only two nonzero components:

$$\ldots, \begin{bmatrix} \cdot \\ 1 \\ 1 \\ 0 \\ 0 \\ \cdot \end{bmatrix}, \begin{bmatrix} \cdot \\ 1 \\ -1 \\ 0 \\ 0 \\ \cdot \end{bmatrix}, \begin{bmatrix} \cdot \\ 0 \\ 0 \\ 1 \\ 1 \\ \cdot \end{bmatrix}, \begin{bmatrix} \cdot \\ 0 \\ 0 \\ 1 \\ -1 \\ \cdot \end{bmatrix}, \ldots$$

This is an orthogonal basis because the vectors are mutually perpendicular. It is an orthonormal basis when divided by $\sqrt{2}$.

The heart of the book is the construction of other (and better) filter banks. The filters will be longer and they will overlap. Orthogonality will not be automatic and it may not be true. The bases are *the impulse responses of the filters* — with a double shift as above. Here is an example with four nonzero components:

$$\ldots,\begin{bmatrix}\cdot\\1\\3\\3\\1\\0\\0\\\cdot\end{bmatrix},\begin{bmatrix}\cdot\\1\\3\\-3\\-1\\0\\0\\\cdot\end{bmatrix},\begin{bmatrix}\cdot\\0\\0\\1\\3\\3\\1\\\cdot\end{bmatrix},\begin{bmatrix}\cdot\\0\\0\\1\\3\\-3\\-1\\\cdot\end{bmatrix},\ldots$$

Those basis vectors are *not* orthogonal. We are alternating lowpass filters with highpass filters, but the first and third (lowpass and shifted lowpass) are not orthogonal. Remember that these vectors are the columns of T^{-1}. In our other language ***they are the impulse responses from the synthesis filters***. The synthesis filters combine to reconstruct the signal, which is exactly what the basis vectors do.

In this nonorthogonal case, the inverse is not the transpose. The basis is not self-orthogonal. It is *biorthogonal* to a different basis — which we now discuss.

Biorthogonal Bases (Dual Bases)

The basis $\{r_n\}$ is biorthogonal to the basis $\{v_n\}$ if the inner products are

$$\left\langle r_i, v_j \right\rangle = \delta(i-j). \tag{2.47}$$

This is the same property that governs the rows of a matrix T and the columns of T^{-1}. It comes directly from $TT^{-1} = I$. So when the columns of T^{-1} are a basis (as above), the rows of T are the biorthogonal basis — the *dual basis.*

We transpose T to turn its rows into columns. We transpose $T^{-1}T = I$ into $T^{*}T^{-*} = I$. Then the same idea that gave the basis $\{v_n\}$ from T^{-1} now gives the biorthogonal basis $\{r_n\}$ from the columns of the transpose matrix T^{*}:

$$v = T^{*}T^{-*}v = \sum_{n=1}^{N} (\text{column } n \text{ of } T^{*})(\text{row } n \text{ of } T^{-*})v$$

The nth dual basis vector is $r_n = \text{column } n \text{ of } T^{*}$

The nth dual coordinate is $d_n = (\text{row } n \text{ of } T^{-*})v = \langle v_n, v\rangle$.

When $T^{*} = T^{-1}$, the basis is self-dual and $v_n = r_n$. We have an orthonormal basis. In general the rows of T and columns of T^{-1} produce biorthogonal bases — provided always that T and T^{-1} are bounded. In this case we have one expansion of v coming from $T^{-1}T = I$, and another expansion from $(T^{-1}T)^{*} = I$:

Theorem 2.4 *If r_n and v_n are biorthogonal bases then any v has two expansions:*

$$v = \sum c_n v_n = \sum \langle r_n, v\rangle v_n \quad \textit{and} \quad v = \sum d_n r_n = \sum \langle v_n, v\rangle r_n. \tag{2.48}$$

In other words $\sum v_n r_n^T = I$ and also $\sum r_n v_n^T = I$.

Example 2.10. The four-tap filters in the example above are biorthogonal to these four-tap filters (after we divide by 16). Check the inner products:

$$\dots,\begin{bmatrix}\cdot\\-1\\3\\3\\-1\\0\\0\\\cdot\end{bmatrix},\begin{bmatrix}\cdot\\-1\\3\\-3\\1\\0\\0\\\cdot\end{bmatrix},\begin{bmatrix}\cdot\\0\\0\\-1\\3\\3\\-1\\\cdot\end{bmatrix},\begin{bmatrix}\cdot\\0\\0\\-1\\3\\-3\\1\\\cdot\end{bmatrix},\dots$$

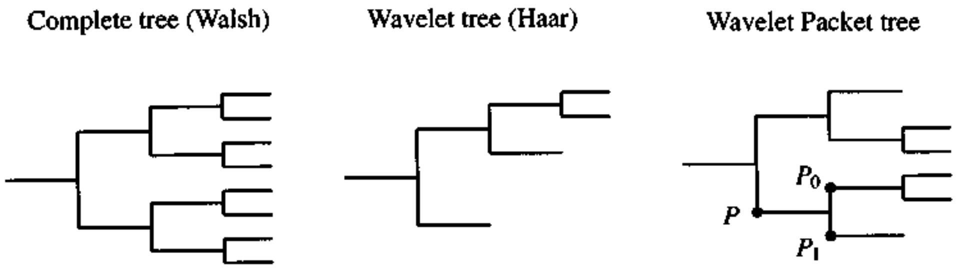
The construction seems magical. We want more like this. Chapters 4 and 5 show how to get more—and the filters need not have equal length. These are the synthesis filters (in T^{-1}) and the analysis filters (in the rows of T) of a *perfect reconstruction filter bank.* The filter bank is biorthogonal.

The norms of T and T^{-1}, and therefore the condition number $\|T\|\,\|T^{-1}\|$, are easily computed for these bases. The double shift in T means that the transform involves a 2×2 matrix function of ω. Maximizing its norm over ω yields $\|T\|$, and maximizing the norm of the 2×2 inverse yields $\|T^{-1}\|$.

A small note of caution. The scaling functions and wavelets will come from iterating the lowpass filter, with rescaling. Not every biorthogonal filter bank leads to biorthogonal bases. Sometimes the iteration diverges. There is a serious step from $L^2(\mathbf{Z})$ in discrete time to $L^2(\mathbf{R})$ in continuous time. The example above fails! Those particular filters with ±1 and ±3 are only good when they are not iterated too often.

Wavelet Packets and the Best Basis

In Chapter 1 only the lowpass filter was iterated. It was assumed that lower frequencies contained more important information than higher frequencies. For many signals this is not true. A *wavelet packet basis allows any dyadic tree structure* (Figure 2.10). At each point in the tree we have the option to send the signal through the lowpass-highpass filter bank, ***or not.***



Complete tree (Walsh)

Wavelet tree (Haar)

Wavelet Packet tree

Figure 2.10: Each wavelet packet tree yields a basis, including Walsh and Haar.

One possibility is the *logarithmic tree,* with lowpass iteration only. Another possibility is the *complete tree,* analogous to the Short Time Fourier Transform. Wavelet packets make up the entire family of bases. Each one is associated with a particular *quadtree,* because it comes from splitting into two (or not splitting) at each step. The decision to split or to merge should be aimed at achieving minimum distortion D—subject to cost and capacity constraints on the rate R.

The main point is that a library of wavelet packet bases is a practical possibility [W]. For a given signal and a given point P in the tree, we have the coefficients of the basis functions $w_P(t)$ for that point. If we choose to split, that set of functions is replaced by two half-size sets of functions—created by lowpass and highpass filtering:

$$\sum_{k=0}^{N} h_0(k) w_P(2t-k) \quad \text{and} \quad \sum_{k=0}^{N} h_1(k) w_P(2t-k).$$

Together with their time shifts, these are the new basis functions for the points P_0 and P_1 in the tree. At each of those new points we again have the option to split.

A point is at level j in the tree if it is reached after j splittings. The basis functions at the root (level $j = 0$) are the shifts $\phi(t-k)$ of the scaling function.

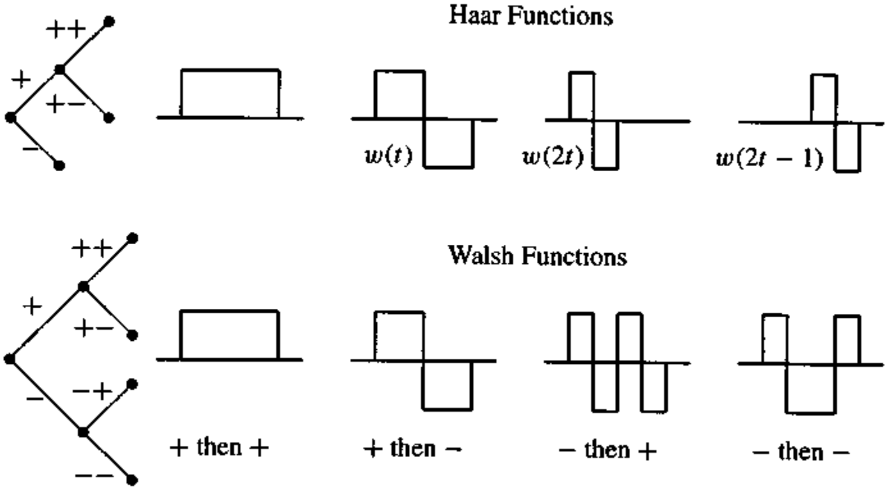


Figure 2.11: Haar iterates only the lowpass filter. Walsh iterates both filters.

Example 2.11. When $\phi(t)$ is the box function, the wavelet $w(t)$ is Haar's up-down square wave. The filter coefficients are $h_0(k) = 1, 1$ for lowpass and $h_1(k) = 1, -1$ for highpass, all divided by $\sqrt{2}$. We describe three special bases for the functions that are piecewise constant on intervals of length 2^{-J}. Within $[0, 1]$ this is a space of dimension 2^J:

1. ***Box basis:*** $\phi(2^J t - k)$ for $0 \le k < 2^J$.
2. ***Haar wavelet basis:*** $\phi(t-k)$ and $w(2^j t - k)$ for $j < J$ and $0 \le k < 2^j$.
3. ***Walsh basis:*** two functions $w_{J-1}(2t) \pm w_{J-1}(2t-1)$ from the basis for $J - 1$.

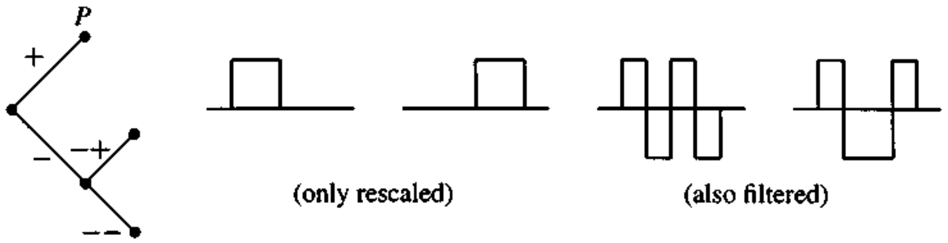
A Walsh basis function takes values 1 *or* − 1 *over the whole unit interval.* It comes from the complete tree with all branches. A wavelet basis function from the logarithmic tree is zero over most of $[0, 1]$. It is nonzero on an interval of length 2^{-j}, varying with j. The box basis functions are nonzero over intervals of fixed length 2^{-J}.

Figure 2.11 shows the four Haar and four Walsh functions (with + or − choices indicated). The *Walsh basis* chooses to split every time. The basis functions are like discrete cosines. *The* *wavelet basis* splits only along one branch. This gives dilation and translation. The *box basis* (scaling function basis) never splits. This gives translation only — a row of small boxes.

The frequency of the ordinary cosine is replaced by the number of *zero-crossings* (sign changes) of the piecewise constant Walsh function. A typical wavelet packet falls between Haar and Walsh. It makes the Walsh decisions, + or −, up to certain points P on the tree. At those points it stops splitting and just rescales. The wavelet packet below could be the best basis for a signal with no low frequencies.

The wavelet packet bases are given by simple recursions (not by simple formulas). They inherit the properties of the filter bank — orthogonality or biorthogonality. The Riesz constants A, B and the condition number B/A are not necessarily in control as J increases. The condition number stays bounded for the logarithmic tree wavelet basis [CoDa]. The condition number is unbounded for the general wavelet packet tree.

We emphasize the recursive form of every wavelet packet decomposition. The coefficient sequence at a point P is treated in exactly the same way as the original sequence $x(n)$ at the original root of the tree. The packet splits or not. If it splits then the two new branches end in two new decision points.



(only rescaled) (also filtered)

There is a corresponding Fourier packet. Its decision is whether to break the interval in two. Instead of a fixed length 8 for all DFT blocks, and 8 × 8 for an image, the splitting decision is made locally — depending on the signal. The "butterfly" in the FFT is like the lowpass-highpass bank for wavelets. But the FFT admits all frequencies 0, 1, 2, 3, ... where the wavelets are dyadic and octave-based.

Now we turn from bases to frames, and give up linear independence.

Frames and Frame Bounds

A frame $\{v_n^+\}$ has one property of a Riesz basis $\{v_n\}$ — every vector or function can be represented as $\sum c_n v_n^+$ with control of $\sum |c_n|^2$. But the requirement of linear independence is dropped. A frame is associated with "oversampling" or "redundancy." There are too many vectors for a basis. We could even repeat the same basis vectors several times — this produces a frame but not an interesting one. More interesting is a set of functions like $\{e^{icnt}\}$, which are a basis for $L^2[0,\pi]$ when $c = 1$ and a tight frame for $c < 1$ (higher than Nyquist rate). For $c = \frac{1}{2}$ this frame is a union of two bases $\{e^{int}\}$ and $\{e^{int}e^{it/2}\}$.

The place to start is in finite dimensions. The M rows of a rectangular matrix give a frame for $\mathbf{R}^N$ — *provided the columns are independent*:

T is any $M \times N$ matrix with N independent columns.

In general $M > N$ and T^{-1} does not exist.

The left-inverse $T^+ = (T^*T)^{-1}T^*$ does exist and $T^+T = I$.

Linear algebra [S] says that T^*T always has the same rank as T. By assumption of independent columns, this rank is N. The $N \times N$ matrix T^*T with this rank is invertible. The identity $(T^*T)^{-1}\, T^*T = I$ shows that $T^+T = I$.

This matrix $T^+ = (T^*T)^{-1}\, T^*$ is a left-inverse and a "pseudo-inverse" of T. The columns of T^+ and the rows of T are two frames, *dual to each other*. The key is $T^+T = I$.

Example 2.12. The three rows r_1, r_2, r_3 of T constitute a frame for $\mathbf{R}^2$:

$$T = \begin{bmatrix} 2 & 0 \\ -1 & \sqrt{3} \\ -1 & -\sqrt{3} \end{bmatrix} \quad \text{has independent columns and} \quad T^*T = \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}.$$

This example is actually a ***tight frame***, because T^*T is a multiple of I. The left-inverse is $(T^*T)^{-1}\, T^* = \frac{1}{6}T^*$. This matrix T^+ is 2×3. It cannot possibly be a right-inverse of T.

The frame vectors are not independent but they span the space. *We can recover v from the columns v_1^+, v_2^+, v_3^+ by using inner products $\langle r_1, v\rangle$ and $\langle r_2, v\rangle$ and $\langle r_3, v\rangle$*. The recovery is based on $T^+T = I$:

$$v = T^+Tv = \sum_{n=1}^{M} (\text{column } n \text{ of } T^+)(\text{row } n \text{ of } T)v$$

The nth analysis frame vector is $r_n = \text{row } n \text{ of } T$

The nth coordinate is $\langle r_n, v\rangle = (\text{row } n \text{ of } T)v$

The nth synthesis frame vector is $v_n^+ = \text{column } n \text{ of } T^+$.

In this example the column vector $v = [1 \;\; 1]'$ has coordinates $\langle r_n, v\rangle = 2$ and $-1+\sqrt{3}$ and $-1-\sqrt{3}$. To synthesize v from the three columns of T^+, multiply the columns by those coordinates:

$$T^+(Tv) = \frac{1}{6}\begin{bmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{bmatrix} \begin{bmatrix} 2 \\ -1+\sqrt{3} \\ -1-\sqrt{3} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The point of the pseudo-inverse T^+ (which is one of many left-inverses) is that the sum of squares of the coordinates is as small as possible. This is the key to frames: *control the sum of squares of coordinates*. Now we define a frame in infinite dimensions.

Definition. *An analysis frame is a set of vectors r_n such that*

$$A\|v\|^2 \quad \le \quad \sum |\langle r_n, v\rangle|^2 \quad \le \quad B\|v\|^2 \quad \textit{for all } v. \tag{2.49}$$

$A > 0$ and $B > 0$ are the "frame bounds." A tight frame has $A = B$.

The frame vectors r_n are not required to be independent. But they span the space! The only vector v orthogonal to every r_n is the zero vector, by (2.49).

The frame operator T transforms v into the sequence of numbers $\langle r_n, v\rangle$. In N dimensions these are M numbers, with $M > N$. In infinite dimensions we have infinitely many numbers, and the key is to recover v from these numbers. The purpose of the frame bounds A and B is to make T^*T and $(T^*T)^{-1}$ bounded operators:

$$\text{Note carefully that} \quad \langle v, T^*Tv\rangle = \langle Tv, Tv\rangle = \sum |\langle r_n, v\rangle|^2. \tag{2.50}$$

Inserting into (2.49) gives $A\|v\|^2 \leq \langle v, T^*Tv\rangle \leq B\|v\|^2$. The operator T^*T is bounded by B. Its inverse is bounded by $1/A$. When we choose the tightest A and B, which we might as well do, the norms of T^*T and its inverse are exactly B and $1/A$. *The frame ratio B/A is the condition number of T^*T.*

The frame bounds ensure that v can be stably recovered from the components $\langle r_n, v\rangle$ of Tv. The recovery operator—the synthesis operator—is the left-inverse $T^+ = (T^*T)^{-1}T^*$:

$$v = T^+Tv = \sum \left(\text{column } n \text{ of } T^+\right)\langle r_n, v\rangle = \sum c_n v_n^+. \tag{2.51}$$

The coordinates are $c_n = \langle r_n, v\rangle$. The synthesis frame vectors v_n^+ are the columns of T^+. When $\{r_n\}$ is a basis, T^+ is T^{-1} and the dual frames are dual bases.

Example 2.13. The three rows of T are an analysis frame (not tight) for $\mathbf{R}^2$:

$$T = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad T^*T = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.$$

Any two rows of T are a basis. The three rows are a frame. The frame bounds are $A = 2$ and $B = 3$. Those are the extreme eigenvalues of T^*T—easy to find in this example. There are infinitely many left-inverses of T, and here are three:

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \quad \text{and} \quad T^+ = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix}.$$

The last one is the best one! It is the pseudoinverse $(T^*T)^{-1}T^*$. You see $\frac{1}{3}$ and $\frac{1}{2}$ from inverting T^*T. The columns of T^+ are the dual frame vectors v_1^+ and v_2^+ and v_3^+.

Why is T^+ best? Because it is the "*smallest*" left-inverse. We are reconstructing any vector v in $\mathbf{R}^2$ from the three components c_1, c_2, c_3 of Tv. In exact arithmetic, the only solution to $Tv = c$ is v—and every left-inverse will find it. In actual arithmetic, and in actual measurements, there are errors in c. We choose the least-squares solution to $Tv = c$. That comes from the normal equations $T^*Tv = T^*c$, whose solution is exactly $v = T^+c$. The synthesis frame reconstructs the truest v.

Frames are associated with *oversampling* and *redundancy*. We meet this in irregular sampling when it is difficult to sample at exactly the right rate. Better to sample too often than too seldom. Oversampling is usually better than undersampling. The interpolating functions $\text{sinc}(t - t_n)$ are a basis for the band-limited space when the sampling rate is exactly right, and they are a frame when we oversample. They cannot reproduce all functions if we undersample.

For regular sampling at the times $t_n = nT$, the perfect rate is the Nyquist rate in Section 2.2. It requires two samples per period—the exact band of frequencies is $|\omega| \leq \pi/T$. In reality we often take more measurements than necessary, and reconstruct the signal by least squares.

The least-squares matrix is exactly T^*T! This is the matrix in the normal equations, when T has too many rows ($M > N$). The coefficient matrix T is $M \times N$ and not invertible. But T^*T is invertible exactly when the rows of T are a frame. Instead of T^{-1} the least-squares solution uses $T^+ = (T^*T)^{-1}T^*$.

Note that a pseudoinverse T^+ is defined for all matrices, of arbitrary rank [S]. T^+ is a left-inverse when the rank is N and the columns of T are independent and the rows of T are a frame. Those are equivalent statements about T in finite dimensions.

Mallat [Mt] identifies two approaches to the reconstruction of the signal $\boldsymbol{v}$. If $\boldsymbol{T}^+$ will be used often, we may compute it explicitly. Its columns are the synthesis frame vectors $\boldsymbol{v}_n^+$. For limited use we do not want the inverse of $\boldsymbol{T}^*\boldsymbol{T}$, only the solutions to specific equations with this coefficient matrix. Just as in linear equations $Ax = b$, we seldom want A^{-1} and we more often want $x = A^{-1}b$.

The iterative solution of the reconstruction problem for irregular sampling has been studied by Grochenig. This is an important application. His algorithms are much better than the simplest iterations.

We mention *Kadec's $\frac{1}{4}$ Theorem* for $L^2[0, 1]$: the exponentials $\left\{e^{2\pi i c_n t}\right\}$ are a basis if $|c_n - n|$ stays below a number $L < \frac{1}{4}$. This is nearly a regular sampling.

In our applications, the "rows" of $\boldsymbol{T}$ are usually functions $r_n(t)$. Then $\boldsymbol{T}$ maps functions $f(t)$ in L^2 to sequences $c_n = \langle r_n, f\rangle$ in ℓ^2. The matrix $\boldsymbol{TT}^*$ contains the inner products of those rows with themselves. It is the "Gram matrix" whose (i, j) entry is the inner product $\langle r_i, r_j\rangle$. *The Riesz bounds A and B come from this matrix.* Its norm is B and the norm of its inverse is $\frac{1}{A}$.

Those numbers measure the linear independence of the functions $r_n(t)$. A case of special interest is when the rows $r_n(t)$ are translates $\phi(t - n)$ of a single scaling function. Theorem 6.6 will give a formula for A and B in this case, involving $\widehat{\phi}(\omega)$.

Summary: The bounds on $\boldsymbol{TT}^*$ are the *Riesz constants* — measuring independence of the rows of $\boldsymbol{T}$. The bounds on $\boldsymbol{T}^*\boldsymbol{T}$ are the *frame bounds*. When $\boldsymbol{T}$ is invertible, these ideas are the same — we have a Riesz basis. When $\boldsymbol{T}$ is not invertible the Riesz lower bound is $A = 0$.

Question: Could a frame include the zero vector $r = 0$?

Answer: Yes. The inner product $\langle 0, f\rangle = 0$ would not harm (or help) the frame bounds. A zero row of $\boldsymbol{T}$ has no effect at all on $\boldsymbol{T}^*\boldsymbol{T}$. And it destroys $\boldsymbol{TT}^*$.

Question: Do a lowpass and highpass filter jointly produce a frame?

Answer: Yes, if their frequency responses have $|C(\omega)| + |D(\omega)| \geq A > 0$.

Question: When would the two filters jointly produce a tight frame?

Answer: When $|C(\omega)|^2 + |D(\omega)|^2 =$ constant. This happens in an orthonormal filter bank. The frame constant is 2! Subsampling removes the redundancy and reduces that constant to 1. The tight frame becomes an orthonormal basis.

Construction of Frames

In finite dimensions, any set of vectors that spans $\mathbf{R}^N$ is a frame. In infinite dimensions, the frame condition is not so simple. *The good constructions start with only one function.* We never want to compute with an infinite set of totally unrelated functions. The natural idea is to create an infinite family from that one function, in a systematic way.

Two systems are of special importance and they lead to ***windows*** and ***wavelets***:

1. ***Windows*** come from one window $g(t)$ by *modulation* and *translation*:

$$g_{mn}(t) = e^{im\Omega t} g(t - nT). \tag{2.52}$$

2. ***Wavelets*** come from one wavelet $w(t)$ by *dilation* and *translation*:

$$w_{jk}(t) = a^{j/2} w\left(a^j t - kT\right). \tag{2.53}$$

The indices m, n, j, and k extend over the set $\boldsymbol{Z}$ of all integers. We have a family $g_{mn}(t)$ of windows and a family $w_{jk}(t)$ of wavelets. The numbers Ω and T and a are fixed. *Those numbers decide whether the family is a frame.* Dividing T by 2 gives twice as many functions — increasing the likelihood of a frame. Dividing Ω by 2 works the same way; so does replacing a by $\sqrt{a}$. The crucial parameter is ΩT for the window family and $T \log a$ for the wavelet family. When these parameters are small, we have more functions and more likely a frame.

As T and Ω and $\log a$ approach zero, we begin to lose in efficiency. The ratio B/A eventually increases — the frame becomes excessively redundant. (Possibly B/A is a convex function of ΩT.) Most of this book is about $a = 2$ and $T = 1$ and special wavelets $w(t)$, leading not only to a frame but to a basis. Section 8.4 is about $\Omega = 1$ and $T = 1$ and special windows, again leading to a basis. Here we are only aiming for frames — much easier.

A point about normalization. The factor $a^{j/2}$ is included in the wavelets to keep the same norm. Without this factor the norms would go to zero as $j \to \infty$. The information to reconstruct $f(t)$ is still available in the inner products with the frame vectors, but the scaling is poor. This emphasizes that in infinite dimensions, the frame must span the space *stably.*

Window Frames: Examples and Theorems

Example 2.1. Suppose $g(t)$ is the unit box function on $[0, 1]$. For $T > 1$ the $g_{mn}(t)$ are not a frame. None of the boxes $g(t - nT)$ overlap the interval $[1, T]$. Any function $f(t)$ supported on this interval has zero inner product with all the windows. These spaced-out windows do not span L^2.

For $T = 1$ the translated boxes $g(t - n)$ fit tightly. Within each box we have exponentials $e^{im\Omega t}$. For $\Omega = 2\pi$ this is an orthonormal basis on the interval $[0, 1]$. For $\Omega < 2\pi$ it is a frame.

We see how a basis appears on the "edge" of a family of frames. As soon as Ω passes 2π, there are not enough exponentials $e^{im\Omega t}$ and the frame is lost. (For $\Omega = 4\pi$ we will completely miss $e^{i2\pi t}$.) In the next example, and often, no basis appears — we go directly from a frame for small ΩT to failure for larger ΩT.

Example 2.2. Suppose $g(t)$ is the Gaussian function $e^{-t^2/2}$ as in [Gabor]. It produces a frame if $\Omega T < 2\pi$. It does *not* produce a frame if $\Omega T = 2\pi$, although this was Gabor's favorite! At the edge of the frames, these functions $g_{mn}(t) = e^{2\pi imt}e^{-(t-n)^2/2}$ do span the space — but not stably. The inner products $\langle f, g_{mn}\rangle$ uniquely determine $f(t)$, but the operator T^*T is not bounded below. The reconstruction formula from those inner products is not numerically stable.

For $\Omega T < 2\pi$ the tables in [D, p. 87] show how B/A depends on Ω and T.

We now state four general results about window frames, without proof. Three are positive, one is negative. The negative one is famous and comes first:

A. *Window frames are impossible with $\Omega T > 2\pi$. Smooth and decaying window frames are impossible with $\Omega T = 2\pi$.*

"Smooth and decaying" in this Balian-Low Theorem [BeHaWa] means that $g'(t)$ and $tg(t)$ are in $L^2(\boldsymbol{R})$. The window bases in Section 8.4 escape this restriction by a simple change in their construction (*cosines replace $e^{2\pi imt}$*). The impossibility for $\Omega T > 2\pi$ first came from Rieffel.

B. *The dual to a window frame comes from a dual window $\widetilde{g}(t)$.*

This makes it worthwhile to compute the dual window. One function $\widetilde{g}(t)$ gives the whole dual frame. In many other cases it is too much work to construct $(T^*T)^{-1}T^*$, whose columns contain the dual frame. Instead we solve a linear system for the synthesis coefficients in $\sum c_n r_n$.

C. *Suppose $\sum |g(t - nT)|^2$ stays between positive constants for all t. Then there is a positive threshold Ω_0 such that $\Omega < \Omega_0$ yields a frame* [D, p. 82].

This assumes that $g(t)$ has compact support or $|t|^{\alpha}|g(t)| \to \infty$ for some $\alpha > 1$. '

D. The frame bounds satisfy $A \le 2\pi \|g\|/\Omega T \le B$.

Wavelet Frames: Examples and Theorems

Example 2.1. The *Mexican hat* $w(t) = (1 - t^2)e^{-t^2/2}$ is the second derivative of the Gaussian. This wavelet is often used in analyzing vision. It has very rapid decay of both w and $\widehat{w}$. For $a = 2$ and small T, the tables in [D, p. 77] show that B/A is very near 1. As T increases we suddenly lose the frame.

Example 2.2. The *Morlet wavelet* is a modulated Gaussian (complex for real signals):

$$w(t) = e^{-i\omega_0 t}e^{-t^2/2} \quad \text{with} \quad \omega_0 = \pi\sqrt{\frac{2}{\ln 2}} = 5.336.$$

This shift in frequency almost gives $\int w(t)\,dt = 0$ as required for wavelets. The actual value is below 10^{-6}, and negligible. The *phase* plot of the wavelet coefficients $\langle f, w_{jk}\rangle$ becomes very useful [Mt] in locating singularities of $f(t)$.

Some applications of these two examples use N different wavelets, called *voices*. This usually produces a tighter frame; B/A comes near 1. A good way to spread the N voices over an octave is by *fractional dilation* of a single wavelet:

$$w_\ell(t) = w\left(2^{-\ell/N}t\right), \quad \ell = 0, \ldots, N-1. \tag{2.54}$$

The negative statement A and positive B in the window rules are reversed for wavelets.

A′. *There are wavelet frames* (even orthonormal bases) *for large values of T $\log a$. The restriction $\Omega T < 2\pi$ on smooth window frames does not apply to wavelets.*

B′. *The dual to a wavelet frame does not always come from one dual wavelet.*

C′. *If $\sum \left|\widehat{w}\left(a^j\omega\right)\right|^2$ stays between positive constants for all ω, and $w(t)$ is smooth, there is a positive threshold such that $T < T_0$ yields a frame* [D, p. 69].

D′. *The frame bounds A and B are related to the wavelet constant C by*

$$A \le \frac{C}{2T\log a} \le B \quad \textit{with} \quad C = 2\pi\int_{-\infty}^{\infty} |\widehat{w}(\omega)|^2\,\frac{d\omega}{|\omega|}. \tag{2.55}$$

For a tight frame equality holds. For an orthonormal basis $A = B = 1$. This constant C will be crucial for the integral wavelet transform in the next section.

Problem Set 2.5

1. Find a formula for the Walsh function that comes from the choices $-+-$. With $J=3$ this function $w_{-+-}(t)$ is equal to 1 or -1 on eight subintervals of $[0, 1]$.

2. Find a general formula for the Walsh functions at level $J=3$. The numbers p, q, r give the three decisions $+1$ or -1.

3. Count how many wavelet packets do not go beyond the level $J=3$.

4. In Example 2.13 with $\boldsymbol{T} = [1\ \ 1\ \ 1;\ \ 1\ \ 0\ \ -1]$ suppose the vector $\boldsymbol{v}$ is $[3\ \ 4]'$.

    1. What are its three inner products c_1, c_2, c_3 with the rows of $\boldsymbol{T}$?
    2. Verify that $2\|\boldsymbol{v}\|^2 \le c_1^2 + c_2^2 + c_3^2 \le 3\|\boldsymbol{v}\|^2$. The frame bounds are $A=2$ and $B=3$.
    3. Verify that $c_1\boldsymbol{v}_1^+ + c_2\boldsymbol{v}_2^+ + c_3\boldsymbol{v}_3^+$ reconstructs this $\boldsymbol{v}$.

5. Suppose the frame vectors are $\boldsymbol{r}_1 = [0\ \ 2]$ and $\boldsymbol{r}_2 = [1\ \ 1]$ and $\boldsymbol{r}_3 = [2\ \ 0]$. Compute $\boldsymbol{T}^*\boldsymbol{T}$ and its eigenvalues A and B (the frame bounds). Also compute $\boldsymbol{T}^+$ and the dual frame.

6. *The bounds for the dual frame* (in the columns of $\boldsymbol{T}^+$) *are* $\frac{1}{B}$ *and* $\frac{1}{A}$. Prove this by showing that $\boldsymbol{T}^+(\boldsymbol{T}^+)^*$ equals $(\boldsymbol{T}^*\boldsymbol{T})^{-1}$. The extreme eigenvalues are ________, and

$$\frac{1}{B}\|\boldsymbol{v}\|^2 \le \sum \left|\langle \boldsymbol{v}_n^+, v\rangle\right|^2 \le \frac{1}{A}\|\boldsymbol{v}\|^2.$$

7. A tight frame has $\boldsymbol{T}^*\boldsymbol{T} = A\boldsymbol{I}$. Explain why this is equivalent to $\sum |\langle \boldsymbol{r}_n, v\rangle|^2 = A\|\boldsymbol{v}\|^2$. The frame bounds A and B are equal. The analysis frame $\{\boldsymbol{r}_n\}$ and synthesis frame $\{\boldsymbol{v}_n^+\}$ are both tight.

8. The Mth roots of 1 are the complex numbers with coordinates $\left(\cos\frac{2\pi k}{M}, \sin\frac{2\pi k}{M}\right)$. Show that these M vectors in $\mathbf{R}^2$ form a tight frame.

9. Prove that $\left\{e^{icnt}\right\}$ is a tight frame for $c < 1$ with constant $A = \frac{1}{c}$.

    Hint: Change $\int_0^{2\pi} v(t)e^{-icnt}\,dt$ to $\int_0^{2\pi} g(s)e^{ins}\,ds$ where $g(s) = \frac{1}{c}v\left(\frac{s}{c}\right)$ up to $s = 2\pi c$ (then zero). The sum of squares of coefficients is $\|g\|^2$ because $\{e^{ins}\}$ is ________ Verify that $\|g\|^2 = \frac{1}{c}\|v\|^2$ so that $A = \frac{1}{c}$.

2.6 Time, Frequency, and Scale

This book emphasizes bases more than frames. The synthesis operators are inverses rather than pseudo-inverses. The analysis bank produces the "right" number of outputs — the data is critically sampled and not oversampled. We want bases that are convenient for computation (by fast transform) and well adapted to the signal (for high compression).

This section is different. First it looks at transforms in continuous time and continuous frequency and continuous scale. Reconstruction is by an integral instead of a sum. This allows a very wide choice of windows and wavelets, in the Short Time Fourier Transform and the Integral Wavelet Transform. The analysis and synthesis formulas are simple and general. But when we sample in time and frequency and scale — in order to select a discrete set as a basis or a frame — the conditions on the windows and wavelets become tighter.

At the end of the section we relax by using many more functions than necessary. Those are ***time-frequency atoms*** — wavelets or windows or whatever. The search for the best representation from a big dictionary of atoms is a very active problem.

May we first contrast windows with wavelets? The competition between them is far from over. Both have the goal of localizing the basis functions. The windowing functions $g(t)$ achieve that by dropping quickly to zero — they can be Gabor windows with the Gaussian factor e^{-t^2}, or they can have finite length. The shifted window $g(t-s)$ is localized around time s. The expansion functions are oscillations inside the window:

$$\textbf{Windowed exponentials} \quad g_{\omega,s}(t) = g(t-s)e^{i\omega t}.$$

These are "time-frequency atoms." When sampled at discrete times $t = n$ and discrete frequencies $\omega = 2\pi k$, the localized exponentials are $g(t-n)e^{2\pi ikt}$. If $g(t)$ is a unit box, these functions give a basis — but not smooth. When $g(t)$ is a smooth window we don't have a basis (Balian-Low Theorem). The samples are over-complete (a frame) or they are incomplete — depending on $g(t)$ and the sampling rate. But Section 8.4 will show how smooth local cosines can give a very satisfactory basis when the frequencies are shifted to $k + \frac{1}{2}$.

Here we keep *all times s and all frequencies ω*. The familiar integral Fourier transform (no window) is a function of ω:

$$\widehat{f}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, e^{-i\omega t} dt. \tag{2.56}$$

The windowed Fourier transform is a function of ω and also the position s:

$$\textbf{Windowed Transform} \quad F(\omega, s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, g(t-s) e^{-i\omega t} dt. \tag{2.57}$$

This is the Fourier transform of the windowed functions $f(t)\,g(t-s)$ for all s. Without the windows, the reconstruction of $f(t)$ from $\widehat{f}(\omega)$ is famous:

$$f(t) = \int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t} d\omega. \tag{2.58}$$

With the windows, we recover $f(t)\,g(t-s)$ by this integral over ω. Equation (2.58) for each s is

$$f(t)\,g(t-s) = \int_{-\infty}^{\infty} F(\omega, s) e^{i\omega t} d\omega. \tag{2.59}$$

Now multiply both sides by $g(t-s)$ (or $\overline{g(t-s)}$ if complex) and integrate over s:

$$f(t) = \frac{1}{\|g\|^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(\omega, s)\, g(t-s) e^{i\omega t}\, d\omega\, ds. \tag{2.60}$$

This is reconstruction from the time-frequency transform $F(\omega, s)$.

Compare this windowed STFT with wavelets, which begin with *one function* $w(t)$. The position variable s still comes from translation to $w(t-s)$. Now comes the difference. *Instead of the frequency variable we have a scale variable.* Instead of modulating the wavelets we rescale them. The "time-scale atoms" are translates and dilates of $w(t)$:

$$\textit{Wavelet functions} \quad w_{a,s}(t) = |a|^{-1/2} w\left(\frac{t-s}{a}\right).$$

The mother wavelet $w(t)$ is $w_{1,0}(t)$ at unit scale $a = 1$ and position $s = 0$. The factor $|a|^{-1/2}$ assures that the rescaled wavelets have equal energy $\|w_{a,s}\| = \|w\|$. We normalize so that all these functions have unit norm $\|w_{a,s}\| = 1$.

Notice how scaling the time by a or a^j automatically scales the translation steps by a^{-1} or a^{-j}:

$$w\left(a^j t - k\right) = w\left(a^j[t - ka^{-j}]\right).$$

The mesh length at level j is scaled down by a^{-j}. The "frequency" is scaled up by a^j. This hyperbolic scaling or dyadic scaling or octave scaling ($a = 2$) is a prime characteristic of wavelet analysis.

The integral wavelet transform is an inner product with wavelets, just as the windowed transform was an inner product with windowed exponentials:

$$\textit{Integral Wavelet Transform} \quad F_w(a,s) = |a|^{-1/2}\int_{-\infty}^{\infty} f(t)\,w\left(\tfrac{t-s}{a}\right) dt. \tag{2.61}$$

The transform F_w is defined on the *time-scale plane*. Again there are two variables, but the scale a has replaced the frequency ω. The subscript in F_w indicates this change.

The wavelet transform $F_w(a,s)$ is redundant, just as the windowed transform $F(\omega, s)$ was redundant. If we select a good wavelet, it will be enough to know F_w at a discrete set of scales and positions. That is the construction to implement digitally. Here we reconstruct $f(t)$ from its over-complete transform $F_w(a,s)$.

Important: The key to scaling is not a but $\log a$. The natural scale is *logarithmic*. The differential is not da but da/a. The frequency window da is proportional to a (this is often called *constant-Q*). Convolution with $\frac{1}{a}w\left(\frac{t}{a}\right)$ gives a unitary operator, when we integrate with respect to logarithmic measure da/a:

$$\int_0^\infty C_a C_a^* \frac{da}{a} = I \quad \text{for convolution } C_a \text{ with } \frac{1}{a}w\left(\frac{t}{a}\right).$$

This is *Calderón's identity*, rediscovered by Grossmann and Morlet. It is proved in Theorem 2.5 below, with a different normalization of the wavelet and wavelet transform (each by $|a|^{-1/2}$). It gives the reconstruction formula that inverts (2.61) and recovers $f(t)$:

Reconstruction from Wavelet Transform

$$f(t) = \frac{1}{C}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_w(a,s)|a|^{-1/2}\,w\left(\frac{t-s}{a}\right)\frac{da\,ds}{a^2}. \tag{2.62}$$

The constant for windows was $C = \|g\|^2$. The constant for wavelets is not $\|w\|^2$, which would be 2π times the integral of $|\widehat{w}|^2$. Instead the constant is $C = 2\pi\int|\widehat{w}|^2 d\omega/|\omega|$. Effectively, C is finite when the transform of the wavelet is zero at $\omega = 0$. This means that *the integral of the wavelet is zero*. Any smooth decaying function $w(t)$ is a mother wavelet for the integral transform provided $\int w(t)\,dt = 0$.

The Haar wavelet was not the box function. It was the up-down square wave with integral zero, one box minus another box. Other wavelets will be combinations of scaling functions $\phi(2t - k)$, *with coefficients that add to zero.* Those coefficients come from the highpass filter! They will be specially chosen, and $\phi(t)$ will be specially chosen, to give a discrete basis. In the continuous time-scale plane we only require that $\int w(t)\,dt = 0$.

The reconstruction formula (2.62) comes from the following theorem.

Theorem 2.5 *For any $f(t)$ and $g(t)$ in L^2, and C as above,*

$$C\int_{-\infty}^{\infty} f(t)\overline{g(t)}\,dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_w(a,s)\,\overline{G_w(a,s)}\,\frac{da\,ds}{|a|^2}. \tag{2.63}$$

With $f = g$ this is a "Parseval formula" in the time-scale plane. The energy in f equals the energy in its wavelet transform, when the time-scale area is measured correctly.

Ironically, the simplest proof of (2.63) uses the Fourier transform. For each a, the transform in (2.61) is a convolution of $f(t)$ with the scaled wavelet $|a|^{-1/2}w\left(\frac{-t}{a}\right)$. The Fourier transform of this convolution F_w is a multiplication $|a|^{1/2}\widehat{f}(\omega)\overline{\widehat{w}(a\omega)}$. Similarly, the transform of $\overline{G}_w$ is a multiplication $|a|^{1/2}\overline{\widehat{g}(\omega)}\widehat{w}(a\omega)$. Then for each a, the integral over s on the right side of (2.63) becomes

$$\int_{-\infty}^{\infty} F_w(a,s)\overline{G_w(a,s)}ds = 2\pi \int_{-\infty}^{\infty} |a|\,\widehat{f}(\omega)\,\overline{\widehat{g}(\omega)}\,|\widehat{w}(a\omega)|^2\,d\omega. \tag{2.64}$$

Integrate this with respect to $da/|a|^2$ and reverse the order of integration. The integral over a gives the predicted constant C, after changing variables to $a\omega$. The right side of (2.63) becomes

$$2\pi \int_{-\infty}^{\infty} \widehat{f}(\omega)\overline{\widehat{g}(\omega)} \int_{-\infty}^{\infty} |\widehat{w}(a\omega)|^2\,da\,d\omega = 2\pi C \int_{-\infty}^{\infty} \widehat{f}(\omega)\overline{\widehat{g}(\omega)}\,d\omega. \tag{2.65}$$

This equals the left side of (2.63) and completes the proof.

Suppose $g(t) = \delta(t)$ is the Dirac delta function (not in L^2). Then equation (2.63) formally becomes the reconstruction formula (2.62) at $t = 0$. The wavelet transform $G_w(a,s)$ of $\delta(t)$ is the wavelet $|a|^{-1/2}w\left(-\frac{s}{a}\right)$. This appears in (2.62) when $t = 0$. By shifting the delta function we reach the reconstruction formula at all t.

This use of the delta function is more legal and familiar than it seems. The ordinary Fourier reconstruction in (2.58) is reached the same way. The analogue of (2.63) is Parseval's equation

$$\int_{-\infty}^{\infty} f(t)\overline{g(t)}\,dt = 2\pi \int_{-\infty}^{\infty} \widehat{f}(\omega)\overline{\widehat{g}(\omega)}\,d\omega \quad \text{for all } f, g \text{ in } L^2. \tag{2.66}$$

With $g(t) = \delta(t)$ this is the reconstruction $f(0) = \int \widehat{f}(\omega)\,d\omega$ at $t = 0$. A shifted delta function gives the inversion formula at any other t. These integrals are correct *at each t* when f is smooth. They are correct *in the L^2 sense* for a general function f in L^2. Technically, we smooth f by restricting its transform to $|\omega| \leq \Omega$ and then let $\Omega \to \infty$.

Note. For real wavelets $w(t)$, when $|\widehat{w}|^2$ is an even function, integrate only over $a > 0$ and compensate by taking only half of the constant C:

$$C \to \int_0^{\infty} |\widehat{w}(\omega)|^2 \frac{d\omega}{\omega} = \tfrac{1}{2}C.$$

The Wigner-Ville Transform

The time-frequency analysis of a signal goes much further than windows and wavelets. By restricting to the choice of one function $g(t)$ or $w(t)$, we create good algorithms — which is our principal purpose. By allowing more general transforms, we can hope to get precise information about the "instantaneous frequency" and "instantaneous spectrum" of a signal. The windows and wavelets do a little smearing — not too much, and the time-frequency trace of the signal can be followed. But Wigner and Ville and many others wanted a sharp trace.

We shall devote just a very short space to this basic topic — the correlation of a signal with itself. Everywhere else we are correlating the signal with chosen windows or wavelets. As those move and rescale, we identify our signal from the correlations. The self-correlation gives an energy density that should automatically pick out the position and frequency and scale of the signal.

We define the *Wigner-Ville transform* of a finite energy function $f(t)$:

$$W(t,\omega) = \int_{-\infty}^{\infty} f\Big(t+\frac{\tau}{2}\Big)\overline{f\Big(t-\frac{\tau}{2}\Big)}e^{-i\omega\tau}\,d\tau. \tag{2.67}$$

Notice especially that W is quadratic in f. We expect $W(t,\omega)$ to be an "energy density"— like the *square* of the window and wavelet transforms, but better. This density has several desirable properties:

$$\frac{1}{2\pi}\int W(t,\omega)\,d\omega = |f(t)|^2 \quad \text{and} \quad \int W(t,\omega)\,dt = \big|\widehat{f}(\omega)\big|^2. \tag{2.68}$$

$W(t-T,\omega-\Omega)$ is the transform of $e^{i\Omega t}f(t-T)$.

$W\left(\frac{t}{a},a\omega\right)$ is the transform of $\frac{1}{a}f\left(\frac{t}{a}\right)$.

$W(t,\omega)$ determines $f(t)$ up to a constant multiplier $|c|=1$.

The transform goes from one variable to two variables. As with wavelets and windows, most functions of two variables are *not* transforms of any $f(t)$. The transform of $\widehat{f}$ flips the variables to reach $W(\omega,-t)$. And the Gaussian plays a special role. We get $W=e^{-t^2-\omega^2}$ from the unit Gaussian and we get modulations of $W=e^{-p^2t^2-q^2\omega^2\pm pqt\omega}$ from all quadratic "chirps."

What good properties does this transform *not* have? First, it is not always positive. Second, the sum of two Gaussians (modulated and shifted) has a transform with *four terms*. Two terms are in the expected positions (T_1,Ω_1) and (T_2,Ω_2), as desired from the Gaussians. Two other terms are in the wrong positions (T_1,Ω_2) and (T_2,Ω_1). These *ghosts* are highly oscillatory.

"Analytic signals" have no negative frequencies: $\widehat{f}(\omega)=0$ for $\omega<0$. Restricted to these signals, the transform is a success. Then $W(t,\omega)$ earns the name "instantaneous spectrum" of $f(t)$. And the instantaneous frequency — *the derivative of the phase* $\phi(t)$ — equals the average $\frac{1}{2\pi}\int \omega W(t,\omega)\,d\omega$ for a large class of signals.

Summary: The integral transforms, by windows and wavelets, make minimal demands on the functions $g(t)$ and $w(t)$. We can analyze and synthesize (transform and inverse transform) with extremely general functions. Reality sets in when the transform becomes discrete — the shifts are multiples of T, the modulations are multiples of Ω, the scales are powers of a. The elegant books by Meyer give an overview of this whole picture.

The rest of this text deals with the discrete transform. We will have very strict conditions on the wavelets! The scales are powers of a fixed integer $a=M$, most often $a=M=2$. Extra conditions are imposed on the wavelet. There are absolute requirements and optional properties. The requirements make it work, the extra conditions make it work well. The goal is to achieve these properties with a fast algorithm:

1. Two-scale or M-scale equation (with special coefficients)
2. Smoothness of the wavelet (optional but desirable)
3. Symmetry or antisymmetry (optional but very desirable)
4. Vanishing moments (one required, more desirable).

Atomic Decompositions

Whether they are wavelets or sinusoids or cosine packets or Gabor functions, we are approximating $f(t)$ by "atoms". A collection of atoms is a "dictionary". We need a reasonable algorithm to choose the nearly best M atoms for a given $f(t)$, out of a dictionary of P atoms. A single basis may be too inflexible, and *the algorithm must adapt to the signal.* Major effort is going into algorithms *not* based on one orthonormal basis.

We will only mention three ideas and their developers. The first is **matching pursuit** (S. Mallat and Z. Zhang). The M atoms are chosen one at a time. The choice at step $k+1$ is the atom that comes closest to the current difference $f_k(t) = f(t) - c_1\phi_1(t) - \cdots - c_k\phi_k(t)$. In practice matching pursuit takes (normalized) inner products $\langle f_k(t), \phi(t)\rangle$ and chooses a large one — an atom $\phi_{k+1}(t)$ that is highly correlated with $f_k(t)$. This is a *greedy and sub-optimal* selection process — greedy because each choice is ignorant of the later choices, sub-optimal because we don't insist on maximizing $\langle f_k(t), \phi(t)\rangle$. The complexity is $O(N \ln N)$ at each iteration, for a signal of length N. The pursuit ends at $k = M$. We can then compute the best combination of the M choices (this is *back-projection*). Experimentally the error approaches white noise as $M \to \infty$.

A second adaptive method is the **best basis** algorithm (R. Coifman and V. Wickerhauser). This begins with a dictionary of bases, often orthonormal. The algorithm chooses the best basis to represent $f(t)$. For wavelet packets from a family of binary trees, the method is particularly well adapted. Available software includes the Wavelet Packet Laboratory for Windows (AK Peters, Wellesley MA 02181-5910). A modification that is optimal in the rate-distortion sense, for compression, is described in [VK, p. 426].

With orthonormal bases at every step, rates and mean-square distortions of the two branches are additive. This allows a fast algorithm to choose the packet that is best adapted to the particular signal. The reader understands that an actual implementation leads to many options. The mean-square ℓ^2 norm may be replaced by other norms (losing additivity at each split but gaining in perceptual quality).

A third method is **basis pursuit** (D. Donoho). The dictionary is still overcomplete. The synthesis $f(t) = \sum c_i\phi_i(t)$ (modelled by $f = \Phi c$) is underdetermined. *Frame theory chooses* **c** *to have a small l^2 norm* (sum of c_i^2 is minimized, leading to linear equations and generalized inverses). *Basis pursuit chooses* **c** *to have a small l^1 norm* (sum of $|c_i|$ is minimized, leading to nonlinear equations and linear programming). This minimizer is generally much sparser — fewer c_i are nonzero.

The l^1 minimizer is also more expensive to compute. In place of the simplex method, interior point and log barrier algorithms associated with Karmarkar solve a sequence of weighted l^2 problems. This can give reasonable success for large dictionaries ($P = 10^4$ and $N = 10^3$). The method can distinguish two nearby bumps in $f(t)$, where matching pursuit will select one centered bump that has high correlation. Linear programming pursues the best basis — *which depends on the signal $f(t)$*. Donoho also studies "empirical atomic decomposition" for $P >> N$. This algorithm strongly controls the number M of atoms $\phi_k(t)$ in the approximation, by minimizing $\|f(t) - \sum c_i\phi_i(t)\|^2 + \lambda M$. The key is in the selection of $\lambda = \sigma\sqrt{2\log P}$. Software is on the web at `http://playfair.stanford.edu`.

The multiplier λ becomes $\sigma\sqrt{2\log N}$ for de-noising with an orthonormal basis. The noise is assumed Gaussian with standard deviation σ. The recommended solution is soft thresholding of the inner products $y_i = \langle \phi_i, f \rangle$ with threshold λ. Each scalar y is shifted toward zero but not beyond:

$$\textbf{\textit{Thresholding}} \quad c_{soft} = (y-\lambda)_+ \quad \text{for} \quad y > 0 \quad \text{and} \quad (y+\lambda)_- \quad \text{for} \quad y < 0.$$

Donoho notes that c_{soft} minimizes $\frac{1}{2}(y-c)^2 + \lambda|c|$. That $|c|$ is the l^1 norm again.

Finally we emphasize that the construction of wavelets has not ended. While writing the book we have learned of *directional* and *translation-invariant* and *steerable* wavelets. And the world of atoms is by no means restricted to wavelets!

Problem Set 2.6

1. Find the Wigner-Ville transform $W(t,\omega)$ when $f(t)$ is the Dirac function $\delta(t)$.
2. Show how $W(t,\omega)$ can also be expressed by an integral of $\widehat{f}$:

$$W(s,\xi) = \frac{1}{2\pi}\int \widehat{f}\left(\xi - \frac{\omega}{2}\right)\widehat{f}\left(\xi + \frac{\omega}{2}\right)e^{is\omega}\,d\omega.$$