LIDS-P-2252
May 1994

Wavelets: A Conceptual Overview

Mitchell M. Livstone
Area I Examination Report
May 1994

Abstract

This report contains a brief conceptual introduction to the theory of wavelets. The basic concepts are developed starting from the Windowed Fourier Transform and time-frequency localization ideas. The continuous (in time and scale) wavelet transform is briefly introduced and is followed by a more detailed discussion of the discrete case. Wavelet frames are defined and the expansion and synthesis equations are developed for redundant discrete wavelet frames. Some of the more important results on discrete wavelet frames are presented next. These include results on tight and orthonormal frames. Finally, it is shown how wavelet theory and multiresolution ideas can be applied to learning systems. In particular, it is shown that by constructing a neural network which has a multiresolution structure of a discrete wavelet frame representation, there exist very simple parameter adjustment rules which imply convergence of the network as well as having other desirable properties related to spatially localized learning.

Contents

1 Introduction

The modern theory of wavelets began to emerge in the early 1980's, however, similar ideas can be traced back to the work of Haar (1910) and Gabor (1946). The field spans many areas in which parts of this theory have been developed independently over the last thirty years. Consequently, an important contribution of wavelets has been the unification of these ideas into a single mathematical theory. Some of this theory was developed in harmonic analysis by Calderon (1964), in quantum mechanics by Aslaksen and Klauder (1968) and in signal processing by Esteban and Galland (1977).

The recent explosion of research in this field has resulted in a unification of theories in different areas as well as the introduction of numerous new ideas. Hence, it is not possible to perform a thorough and brief survey of the entire field. The purpose of this report is to provide the basic conceptual foundation for wavelet theory and present a few of the many important results in this field.

The development in this report follows the path of least conceptual resistance. The starting point for the development of wavelets will be the time-frequency localization idea of the Windowed Fourier Transform (WFT). Going from the WFT to the Continuous Wavelet Transform (CWT) is conceptually trivial. The technical results here deal with admissibility of basis functions, the synthesis formula, and characterizability of functions by the CWT. The step from CWT to the Discrete Wavelet Transform (DWT) is also trivial. However, many interesting technical issues arise in this case. These include characterization and reconstruction of functions, admissibility of the wavelet basis functions, redundancy in the representation, existence of orthonormal bases having infinite or compact support, and many more. Some of these are covered in detail and an attempt is made to provide general comments on the important results and issues not covered in this report.

The main contribution of wavelets is their applicability to a wide variety of problems with a common characteristic. That is, the functions (or signals) of interest contain short duration – high frequency components as well as longer duration – low frequency components. In this case, the wavelet basis can lead to a more parsimonious representation than the Fourier ba-

sis, for example. Proponents of wavelet theory argue that this holds true in many situations. In fact, it can be shown that the human auditory system also processes signals in this way.

Following this theme, it is shown that this theory can be readily applied to the field of neural networks. The network approximation properties and choice of activation functions are given by the wavelet theory. The network then tries to learn the DWT coefficients (weights) incrementally as it receives more samples of the unknown function. It is shown that a simple Least Mean Square (LMS) algorithm can be used to learn the weights. Furthermore, it is shown that the restrictions on the basis functions automatically imply a very desirable property of the neural network. Namely, the low (spatial) resolution neurons are updated more slowly than the high resolution neurons achieving a good trade–off between spatial localization and generalization.

2 The Basic Idea

The standard Fourier representation of signals works well only if the spectral properties of signals are fairly stationary. If the spectral properties change during the time frame of interest, the Fourier Transform is not a good way of characterizing them. The idea of the Windowed Fourier Transform (WFT) is to window the signal in time and perform a Fourier decomposition on the windowed signal as the window slides along the time axis. More precisely, one defines the continuous WFT as

$$(\mathcal{F}^{\text{win}}x)(\omega,\tau) = \int dt\ x(t)g(t-\tau)e^{-i\omega t} \tag{1}$$

where the window function, g and its Fourier transform, $\hat{g}$ are both concentrated around zero. The transform $(\mathcal{F}^{\text{win}}x)(\omega,\tau)$ essentially gives the content of f near time τ and frequency ω.

It is instructive to interpret the WFT in the following two ways [1].The first interpretation of the CWFT is the obvious one of a Fourier Transform windowed in time, $(\mathcal{F}^{\text{win}}x)_\tau(\omega)$. The second is a modulated filter bank, $(\mathcal{F}^{\text{win}}x)_\omega(\tau)$, where the filter impulse response is given by $g_\omega(t) = g(-t)e^{-i\omega t}$. In connection to the first interpretation, one can consider the

ability of the WFT to discriminate two pulses in time. The time resolution of g is defined as

$$\Delta t = \left(\frac{\int t^2 |g(t)|^2 dt}{\int |g(t)|^2 dt} \right)^{1/2} \tag{2}$$

One can also consider the ability to discriminate between two sinusoids. The frequency resolution is defined as

$$\Delta \omega = \left(\frac{\int \omega^2 |\hat{g}(\omega)|^2 d\omega}{\int |\hat{g}(\omega)|^2 d\omega} \right)^{1/2} \tag{3}$$

This means that two pulses that are more than Δt apart in time and two sinusoids that are more than $\Delta \omega$ apart in frequency can be resolved fairly well. There is, however, a fundamental limitation to how small both can become. This limitation is given by the relationship

$$\Delta t \cdot \Delta \omega \geq \frac{1}{4\pi} \tag{4}$$

with equality achieved for Gaussian windows.

The limitation of the WFT is that the window, g, is fixed and therefore so are Δt and $\Delta \omega$. However, in applications where high frequency phenomena have shorter duration than the low frequency components, one would like to have better time resolution at higher frequencies.

What wavelets do is vary Δt and $\Delta \omega$ in a special way [2]. At low frequencies, the wavelets have better frequency resolution and at high frequencies they have better time resolution. In particular, the relationship for wavelets is $\Delta \omega / \omega$ =constant. This can be seen by examining the Continuous Wavelet Transform (CWT) [2].

$$(T^{\text{wav}} f)(a, b) = |a|^{-1/2} \int dt \; f(t) \psi(\frac{t-b}{a}) \tag{5}$$

The translation variable is b (like τ in WFT), the scale a is like $1/\omega$, and ψ is a bandpass function satisfying some additional constraints which will be discussed shortly. Note that as $|a^{-1}|$ increases, the support of $\psi(\frac{t-b}{a})$ decreases which is exactly the goal. The function ψ is called the "mother" wavelet and the rest of the wavelet basis is obtained via dilations and translations of ψ.

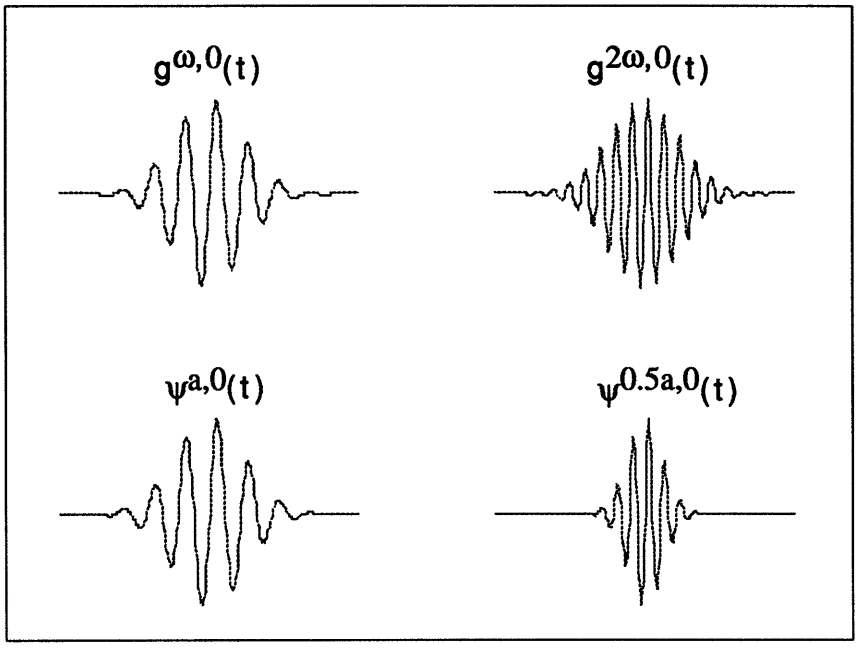

Figure 1: Comparison of WFT to DWT basis functions

The definitions $\psi^{a,b}(t) \equiv |a|^{-1/2}\psi(\frac{t-b}{a})$ and $(T^{\text{wav}}f)(a,b) = \langle f, \psi^{a,b}\rangle$ will be used in the remainder of this report. Moreover, $\hat{\phi}$ will be used to denote the Fourier transform of ϕ.

The following is a comparison of the WFT with the CWT.

$$\underset{\text{Windowed Fourier}}{g^{\omega,\tau}(t) \equiv g(t-\tau)e^{-i\omega t}} \quad \text{vs.} \quad \underset{\text{Wavelet}}{\psi^{a,b}(t) \equiv |a|^{-1/2}\psi(\tfrac{t-b}{a})}$$

This is shown graphically in Figure 1 where one can see that the envelope of the basis functions for the WFT is constant, while the wavelet functions get "squashed" at higher frequencies. When one gains time resolution at higher frequencies, the frequency resolution must be given up. This is usually not a restriction since one typically does not need fine frequency resolution at very high frequencies.

Just as the discrete windowed Fourier transform (DWFT) is obtained by a discretization of τ and ω, the discrete wavelet transform (DWT) is obtained

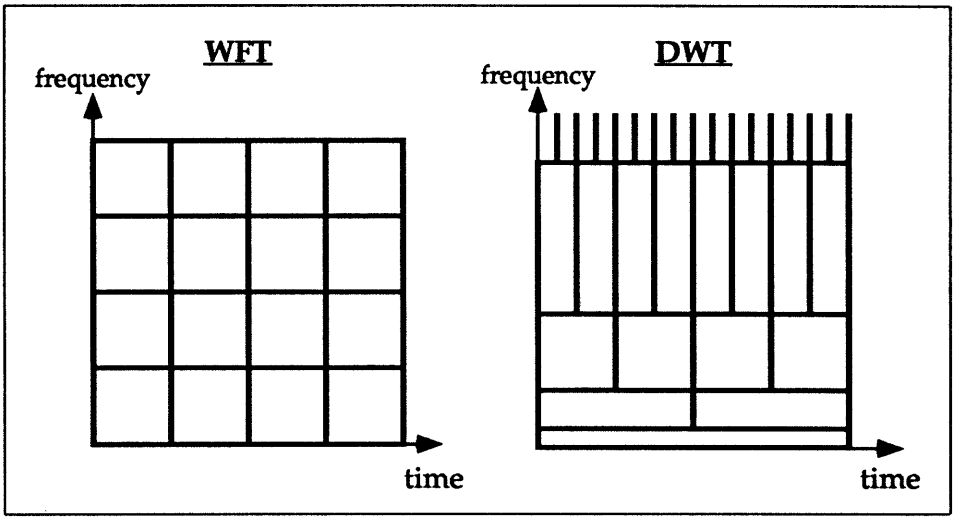

Figure 2: Tiling of the Time–Frequency Plane

by a discretization of the translation variable b and the scale a. More precisely, one chooses $a = a_0^m$ for some fixed $a_0 > 1$. Note that the closer a_0 is to 1, the higher the frequency density (higher redundancy). Because the width of $\psi(a_0^{-m}\cdot)$ is proportional to a_0^m one chooses $b = nb_0a_0^m$ for the translation. Thus, at higher frequencies (smaller m) there are smaller incremental translations (finer time resolution) but larger increments in a (coarser frequency resolution), while at lower frequencies (larger m) there are larger incremental translations (coarser time resolution) but smaller increments in a (finer frequency resolution). This results in the discretized wavelet functions

$$\psi_{m,n} \equiv |a_0|^{-m/2}\psi(a_0^{-m}t - nb_0) \; . \tag{6}$$

Another way to compare the WFT to the wavelet transform is by considering the tiling of the time–frequency plane. Each tile corresponds to a basis function which is essentially time–frequency localized in that particular tile. In the case of the DWFT, the tiling is a regular grid, but in the discrete wavelet case, the time–frequency plane is tiled in a logarithmic manner (see Figure 2). The bottom line is that the WT is better at "zooming" in on short duration, high frequency phenomena.

The following table displays all of the Fourier and wavelet transform formulas for comparison.

Transform	Continuous	Discrete
Fourier	$(\mathcal{F}f)(\omega) = \int dt\ e^{-i\omega t} f(t)$	$(\mathcal{F}f)(n) = \int dt\ e^{-in\omega_0 t} f(t)$
Windowed Fourier	$(\mathcal{F}^{\text{win}}f)(\omega,\tau) = \int dt\ f(t)g(t-\tau)e^{-i\omega t}$	$(\mathcal{F}^{\text{win}}f)_{m,n} = \int dt\ f(t)g(t-n\tau_0)e^{-im\omega_0 t}$
Wavelet	$(T^{\text{wav}}f)(a,b) = \lvert a\rvert^{-1/2}\int dt\ f(t)\psi(\frac{t-b}{a})$	$(T^{\text{wav}}f)_{m,n} = \lvert a_0\rvert^{-m/2}\int dt\ f(t)\psi(a_0^{-m}t - nb_0)$ $a_0 > 1,\ a \sim a_0^m\ \ b \sim nb_0a_0^m$

3 What Can One Study About Wavelets?

The study of wavelets splits naturally into two parts. The first is the continuous wavelet transform which was given in Equation 1. The continuous case is particularly easy because there is an exact reconstruction formula for f in terms of $(T^{\text{wav}}f)(a,b)$ and $\psi^{a,b}$. The second part is the discrete wavelet transform. In this case, there generally does not exist an exact synthesis formula. This leads to two important questions:

1. Do the DWT coefficients $\langle f, \psi_{m,n}\rangle$ completely characterize f? (characterization via $\{\psi_{m,n}\}$)

2. Is it possible to compute f as a linear combinations of the functions $\{\psi_{m,n}\}$? (representation via $\{\psi_{m,n}\}$)

These questions will be answered in the following sections.

Within the above subdivisions, there are many issues which deal with sampling and time–frequency localization properties of wavelets, admissibility conditions for wavelet bases, existence of orthonormal wavelet bases having compact or infinite support (discrete only), regularity of the wavelets

and the wavelets' ability to characterize the regularity of f, as well as many other interesting topics.

4 Continuous Wavelets

This section briefly introduces the continuous wavelet transform, states the synthesis formula and comments on some of the CWT results not covered in this report. Recall the definition of the CWT:

$$(T^{\text{wav}} f)(a,b) = |a|^{-1/2} \int dt\ f(t)\psi(\frac{t-b}{a}) \tag{7}$$

where a is the scale (like $1/\omega$) and b is the translation in time. Thus, the CWT maps $L_2(R)$ into $L_2(R^2)$ (lots of redundancy). It is shown in Daubachies [2] that a function f can be perfectly reconstructed from the CWT via the "*resolution of the identity*" formula:

$$f(t) = C_\psi^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{da\ db}{a^2} \langle f, \psi^{a,b} \rangle \psi^{a,b}(t) \tag{8}$$

$$\text{where } C_\psi = 2\pi \int d\zeta |\hat{\psi}(\zeta)|^2 |\zeta|^{-1} < \infty$$

This implies, in particular, that $\hat{\psi}(0) = 0$ or $\int \psi(t)dt = 0$. In the continuous case these are the only constraints on admissibility of ψ.

4.1 Comments on Other CWT Results

There are numerous results dealing with the CWT. These include general theorems on sampling and reconstruction and are not limited to wavelets. One interesting result shows how the CWT of f can be used to characterize regularity, or Hölder continuity of f.

Definition 4.1 $|f(x) - f(y)| \leq C|x-y|^\alpha$, $\alpha \in (0,1]$ *means* f *is Hölder continuous with exponent* α.

The main result is given by the following theorem which is proved in [2].

Theorem 4.2 *Suppose* $\int dx(1+|x|)|\psi(x)| < \infty$ *and* $\hat{\psi}(0) = 0$. *If* f *is bounded and H–cont with exponent* α, *then* $|(T^{wav} f)(a,b)| \leq C' a^{\alpha+1/2}$

Conversely, suppose that ψ is compactly supported and f is bounded and continuous. If for some $\alpha \in (0,1]$ the CWT of f satisfies $|\langle f, \psi^{a,b} \rangle| \leq C a^{\alpha+1/2}$, then f is H-cont with α.

This says that if a function has a certain amount of regularity, its DWT coefficients will "fall off" with frequency at a certain rate.

Unlike the WFT, the CWT can also characterize *local* regularity of f by considering Hölder continuity at a point. These results are also given in [2].

5 Redundant Discrete Wavelet Frames

This section presents a fairly detailed discussion of redundant discrete wavelet frames. The term redundant means that the wavelets $\{\psi_{m,n}\}$ do not necessarily constitute an orthonormal set (but do span L_2). One of the goals of this section is to show that orthonormality imposes strict constraints on the wavelet functions and it is not even clear until the next section whether there exist any orthonormal families of wavelets besides the Haar basis.

Recall the two questions:

1. Do the DWT coefficients $\langle f, \psi_{m,n} \rangle$ completely characterize f? (characterization via $\{\psi_{m,n}\}$)
2. Is it possible to compute f as a linear combinations of the functions $\{\psi_{m,n}\}$? (representation via $\{\psi_{m,n}\}$)

For the CWT these are both clearly implied by the resolution of identity (ROI) formula. For DWT this is not as clear. Is there a similar ROI formula for DWT and is there a similar admissibility condition for ψ? The answer is "sort of". It will be shown that for reasonable ψ and appropriately chosen a_0 and b_0, there exist $\widetilde{\psi_{m,n}}$ which satisfy

$$f = \sum_{m,n} \langle f, \psi_{m,n} \rangle \widetilde{\psi_{m,n}}$$

It then follows that for any $g \in L_2$

$$\begin{aligned} \langle g, f \rangle &= \sum_{m,n} \langle g, \widetilde{\psi_{m,n}} \rangle \langle \psi_{m,n}, f \rangle \text{ or that} \\ g &= \sum_{m,n} \langle g, \widetilde{\psi_{m,n}} \rangle \psi_{m,n} \text{ in the weak sense.} \end{aligned}$$

This will show that possibly different functions need to be used in the characterization and representation of f.

5.1 The Frame Operator

In order to reconstruct f from $\langle f, \psi_{m,n} \rangle$ one needs

$$\{\langle f_1, \psi_{m,n} \rangle\} \textit{ close to } \{\langle f_2, \psi_{m,n} \rangle\} \Leftrightarrow f_1 \textit{ close to } f_2$$

This really says that the DWT viewed as a linear operator must be bounded and have a bounded inverse. This is examined more rigorously in the following development.

The DWT must be a *bounded* operator from $L_2(R)$ to $l_2(Z^2)$ and will be for any ψ which has some decay in time and frequency, $\hat{\psi}(0) = 0$, $a_0 > 1$, and $b_0 > 0$. The boundedness of the DWT is expressed as

$$\sum_{m,n} |\langle f, \psi_{m,n} \rangle|^2 \leq B \|f\|^2 \, .$$

Using the natural topologies on $L_2(R)$ and $l_2(Z^2)$ to measure closeness, the "only if" part of the above statement becomes:

$$\exists \alpha < \infty \text{ s.t. } \sum_{m,n} |\langle f, \psi_{m,n} \rangle|^2 < 1 \;\Rightarrow\; \|f\|^2 \leq \alpha$$

which means that $A\|f\|^2 \leq \sum_{m,n} |\langle f, \psi_{m,n} \rangle|^2$ for some $A = \alpha^{-1} > 0$.

Putting this together results in the general definition of a frame.

Definition 5.1 *A family $\{\phi_j \; ; \; j \in J\}$ in a Hilbert space $\mathcal{H}$ is called a frame if there exist $A > 0$, $B < \infty$ s.t. for all $f \in \mathcal{H}$*

$$A\|f\|^2 \leq \sum_{j \in J} |\langle f, \phi_j \rangle|^2 \leq B\|f\|^2 \tag{9}$$

The constants A and B are called the frame bounds. If $A = B$, the frame is *tight.* By the use of the polarization identity one can show that in this case

$$f = A^{-1} \sum_j \langle f, \phi_j \rangle \phi_j$$

in the weak sense. Tight frames are not necessarily orthonormal (only if $A = B = \|\phi_j\| = 1$). When the frame is not tight it is necessary to introduce the *frame operator.*

If $\{\phi_j\}$ is a frame in $\mathcal{H}$, the *frame operator* F is the linear operator from $\mathcal{H}$ to $l_2(J)$ defined by

$$(Ff)_j = \langle f, \phi_j \rangle = \overline{\langle \phi_j, f \rangle}$$

The adjoint F^* can easily be computed.

$$\langle F^* c, f \rangle = \langle c, Ff \rangle = \sum_{j \in J} c_j \langle \phi_j, f \rangle$$

$$\text{and so} \quad F^* c = \sum_{j \in J} c_j \phi_j \ .$$

Since Equation 9 implies that $A \leq \|F^* F\| \leq B$, this means that $F^* F$ is invertible and $\|(F^* F)^{-1}\| \leq A^{-1}$.

Let $\tilde{\phi}_j = (F^* F)^{-1} \phi_j$ which results in a new frame. In particular, the family $\{\tilde{\phi}_j\}$ is a frame with bounds

$$B^{-1} \|f\|^2 \leq \sum_{j \in J} |\langle f, \tilde{\phi}_j \rangle|^2 \leq A^{-1} \|f\|^2 \ . \tag{10}$$

The associated frame operator is $\tilde{F} : \ \mathcal{H} \mapsto l_2(J)$ and is defined by

$$(\tilde{F} f)_j = \langle f, \tilde{\phi}_j \rangle \ .$$

It is true that $\tilde{F} = F(F^* F)^{-1}$ and $\tilde{F}^* \tilde{F} = (F^* F)^{-1}$, but the important relation is

$$\tilde{F}^* F = F^* \tilde{F} = \mathcal{I} \ . \tag{11}$$

The new frame $\{\tilde{\phi}_j\}$ is called the *dual frame.* The answer to the characterization and reconstruction questions lies in Equation 11 which really says that

$$\sum_{j\in J}\langle f,\phi_j\rangle\tilde{\phi}_j = f = \sum_{j\in J}\langle f,\tilde{\phi}_j\rangle\phi_j \ . \tag{12}$$

The left side gives a reconstruction formula in terms of DWT coefficients, while the right side gives a superposition representation in terms of the ϕ_j. Thus, the two questions are just duals of one another. An immediate question is how does one actually compute the $\tilde{\phi}_j$'s ? This will be partially answered in the following section.

5.2 Redundancy and Reconstruction

The DWT frames are usually redundant which results in many possible superposition representations of $f = \sum_j c_j\phi_j$. One can show that for any such representation

$$\sum_j |c_j|^2 \geq \sum_j |\langle f,\tilde{\phi}_j\rangle|^2$$

This says that the most "economical" ϕ representation of f is in terms of the coefficients $\langle f,\tilde{\phi}_j\rangle$.

Similarly, there are many possible reconstructions from the DWT. If $f = \sum_j\langle f,\phi_j,\rangle u_j$, one can also show that for any such reconstruction

$$\sum_j |\langle u_j,g\rangle|^2 \geq \sum_j |\langle\tilde{\phi}_j,g\rangle|^2$$

which means that the most economical reconstruction of the DWT of f is via a superposition of $\tilde{\phi}_j$'s.

The reconstruction of f from $\langle f,\phi_j\rangle$ requires the computation of $\tilde{\phi}_j = (F^*F)^{-1}\phi_j$. If $A \approx B$ then $F^*F \approx \frac{A+B}{2}\mathcal{I}$. In fact, the following equality holds.

$$f = \frac{2}{A+B}\sum_{j\in J}\langle f,\phi_j\rangle\phi_j \ + \ Rf$$

where $R = \mathcal{I} - \frac{2}{A+B}F^*F$. It can also be shown that

$$(F^*F)^{-1} = \frac{2}{A+B}(\mathcal{I} - R)^{-1}$$

and $\sum_{k=0}^{\infty} R^k$ converges in norm to $(\mathcal{I} - R)^{-1}$. Now the dual frame is given by

$$\tilde{\phi}_j = \frac{2}{A+B}\sum_{k=0}^{\infty} R^k \phi_j$$

and can be approximated by finite sums which converge exponentially:

$$\|f - \sum_{j\in J}\langle f, \phi_j\rangle \phi_j^N\| \leq \left(\frac{r}{2+r}\right)^{N+1} \|f\|$$

where $\phi_j^N = \frac{2}{A+B}\sum_{k=0}^{N} R^k\phi_j$ and $r = (B/A) - 1$. This gives the motivation for constructing frames that are tight or almost tight.

There exist recursive approximation algorithms for computing $\tilde{\phi}_j$ or f directly, and some of the current research is aimed at speeding up such recursive algorithms [4].

5.3 Admissibility and Frame Bounds

How can one tell if a ψ mother wavelet, a_0 and b_0 will constitute a frame? The first result in this direction gives a necessary condition for admissibility and is proved in [2].

Theorem 5.2 *If $\psi_{m,n}(x) = a_0^{-m/2}\psi(a_0^{-m}x - nb_0)$, $m, n \in Z$ constitute a frame with A, B, then*

$$\frac{b_0 \ln a_0}{2\pi} A \leq \int_0^{\infty} d\xi\ \xi^{-1}|\hat{\psi}(\xi)|^2 \leq \frac{b_0 \ln a_0}{2\pi} B$$

$$\textit{and} \quad \frac{b_0 \ln a_0}{2\pi} A \leq \int_{-\infty}^{0} d\xi\ \xi^{-1}|\hat{\psi}(\xi)|^2 \leq \frac{b_0 \ln a_0}{2\pi} B$$

This places a strong restriction on *tight* frames, especially orthonormal frames. Note that the admissibility condition for the CWT

$$\int d\xi\ \xi^{-1}|\hat{\psi}(\xi)|^2 < \infty$$

falls out of this trivially. The discrete case is indeed much more difficult than its continuous counterpart.

The next result proven in [2] gives a sufficient condition for $\psi_{m,n}$ to constitute a frame.

Theorem 5.3 *If ψ, a_0 satisfy*

$$\inf_{1\le|\xi|\le a_0} \sum_{m=-\infty}^{\infty} |\hat{\psi}(a_0^m\xi)|^2 \;>\; 0$$

$$\sup_{1\le|\xi|\le a_0} \sum_{m=-\infty}^{\infty} |\hat{\psi}(a_0^m\xi)|^2 \;<\; \infty$$

and if $\beta(t) = \sup_\xi \sum_m |\hat{\psi}(a_0^m\xi)|\ |\hat{\psi}(a_0^m\xi + t)|$ decays at least as fast as $(1 + |t|)^{-(1+\epsilon)}$ for some $\epsilon > 0$, then there exists a b_0^* such that $\psi_{m,n}$ constitute a frame for all $b_0 < b_0^*$. The following frame bounds also hold.

$$A = \frac{2\pi}{b_0}\left\{\inf_{1\le|\xi|\le a_0} \sum_{m=-\infty}^{\infty} |\hat{\psi}(a_0^m\xi)|^2 - \sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left[\beta\left(\frac{2\pi}{b_0}k\right)\beta\left(-\frac{2\pi}{b_0}k\right)\right]^{1/2}\right\}$$

$$B = \frac{2\pi}{b_0}\left\{\sup_{1\le|\xi|\le a_0} \sum_{m=-\infty}^{\infty} |\hat{\psi}(a_0^m\xi)|^2 - \sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left[\beta\left(\frac{2\pi}{b_0}k\right)\beta\left(-\frac{2\pi}{b_0}k\right)\right]^{1/2}\right\}$$

The main point of all this is that for many different ψ's, there are ranges of a_0, b_0 such that $\psi_{m,n}$ constitute a frame. The challenge is in trying to construct tight or even orthonormal frames. A typical construction of a tight frame will be shown later. Then, in Section 6, a systematic procedure for constructing orthonormal frames will be presented.

5.3.1 Connection to WFT Frames

The preceding analysis on wavelet frames is not limited to wavelets and the windowed Fourier transform can be analyzed in the same way. The discretization of $g_{\omega,t}(x) = e^{i\omega x}g(x-t)$ is

$$g_{m,n}(x) = e^{im\omega_0 x}g(x - nt_0)$$

One can analyze cases when $g_{m,n}$ constitute a frame and look at expansion and reconstruction formulae. One result is

$$A \leq \frac{2\pi}{\omega_0 t_0}\|g\|^2 \leq B$$

This implies that for an orthonormal basis (assuming g is normalized), $g_{m,n}$ constitute a frame only if $\omega_0 t_0 = 2\pi$. This is in contrast to the wavelet case where there is no such constraint on a_0, b_0. One can derive similar results for frame bounds and dual frames, and construct window functions which result in tight frames.

The main point is that all of the wavelet analysis is not unique to wavelets since the wavelet is only *one* way of tiling the time–frequency plane. There are many possible ways of doing this and some of the current research deals with the construction of orthonormal frames having more general and even time varying tilings of the time–frequency plane [5].

5.4 Construction of Tight Frames

This section briefly shows a construction of a tight frame. The idea is to construct a $\hat{\psi}$ having finite support and satisfying the equality for the frame bound. The first step is to define some C^k function, ν such that $\nu(x) = 0$ for $x \leq 0$ and $\nu(x) = 1$ for $x \geq 1$. One example is the following function.

$$\nu(x) = \begin{cases} 0 & x \leq 0 \\ \sin^2 \frac{\pi}{2} x & 0 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

For arbitrary $a_0 > 1$ and $b_0 > 0$ define $\hat{\psi}$ as follows.

$$\hat{\psi}(\xi) = [\ln a_0]^{-1/2} \begin{cases} 0 & \xi \leq l \text{ or } \xi \geq a_0^2 l \\ \sin\left[\frac{\pi}{2}\nu\left(\frac{\xi - l}{l(a_0-1)}\right)\right] & l \leq \xi \leq a_0 l \\ \cos\left[\frac{\pi}{2}\nu\left(\frac{\xi - a_0 l}{a_0 l(a_0-1)}\right)\right] & a_0 l \leq \xi \leq a_0^2 l \end{cases}$$

where $l = 2\pi[b_0(a_0^2 - 1)]^{-1}$. This function satisfies

$$\text{support}(\hat{\psi}) = l(a_0^2 - 1) = 2\pi/b_0 \text{ and}$$
$$\sum_{m \in Z} |\hat{\psi}(a_0^m \xi)|^2 = (\ln a_0)^{-1} \chi_{(0,\infty)}(\xi)$$

where χ is the indicator function. One can then show that for any $f \in L_2$

$$\sum_{m,n \in Z} |\langle f, \psi_{m,n} \rangle|^2 = \frac{2\pi}{b_0 \ln a_0} \|f\|^2$$

which shows that the frame is indeed tight.

5.5 Truncated Reconstructions

In practice, only a finite number of basis functions can be used in the computations. One can derive results which essentially say the following. If the mother wavelet is sufficiently time–scale localized and if f is essentially globally localized in some region of time and frequency, then only a finite number of $\psi_{m,n}$'s are necessary to approximate f very well. Again, this analysis can be performed on wavelet frames as well as windowed Fourier frames. This is an important topic for actual implementation but we will not dwell on it here.

5.6 What Does Redundancy Buy?

There is evidence that computations are more robust to errors in coefficients. This can be explained by noting that the more redundant the frame, the "smaller" is Ran(F). This means that if there are random errors in the DWT coefficients (e.g., quantization error), more of this error will be perpendicular to Ran(F) and hence, will not contribute to the reconstruction of f. One can construct simple finite dimensional examples to illustrate this fact [2], but general results are still unsolved (at least in 1992).

6 Existence of Orthonormal Wavelet Bases

One of the most important results in wavelet theory is the existence of orthonormal bases which, unlike the Haar and Littlewood–Paley bases, have good time and frequency localization. This is an important improvement over the WFT for which there is a theorem which says that given an orthonormal WFT frame, one must sacrifice either good time or frequency localization. One can use a construction similar to the tight frame construction with some additional tricks. However, the next section discusses a remarkable innovation which allows one to systematically construct orthonormal wavelet bases.

6.1 Multiresolution Analysis

The technique of multiresolution analysis was developed by Mallat and Meyer in 1986. This explained some of the "magic" behind the constructions of orthonormal wavelet bases. Multiresolution analysis uses a nested sequence of subspaces in L_2 which are just scaled (by two) versions of each other and tend to all of L_2. In particular, let the closed subspaces V_j satisfy

$$\cdots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots$$

$$\begin{aligned} \overline{\bigcup_{j \in Z} V_j} &= L_2(R) \\ \bigcap_{j \in Z} V_j &= \{0\} \end{aligned}$$

Two other requirements for multiresolution analysis is scaling by two and shift invariance. More precisely,

$$f \in V_j \Leftrightarrow f(2^j \cdot) \in V_0 \text{ and}$$

$$f \in V_0 \Rightarrow f(\cdot - n) \in V_0 \text{ for all } n \in Z$$

Note that the two requirements imply that

$$f \in V_j \Rightarrow f(\cdot - 2^j n) \in V_j \text{ for all } n \in Z \ .$$

One example of spaces satisfying both requirements is

$$V_j = \{f \in L_2 \ : \ f|_{[2^j k, 2^j(k+1)]} = \text{constant}\}$$

This is called the Haar multiresolution analysis. One final requirement is the existence of a ϕ such that $\phi_{0,n}$ is an orthonormal basis for V_0. This implies that for a fixed j, $\phi_{j,n}$ is an orthonormal basis for V_j. The Haar example is shown in Figure 3.

Note that $\{\phi_{j,n} \ : \ j,n \in Z\}$ *is not an orthonormal basis for* L_2 and there is no cheating going on. One starts with ϕ_n, an orthonormal basis for V_0 and gets $\psi_{m,n}$, an orthonormal basis for *all of* L_2!

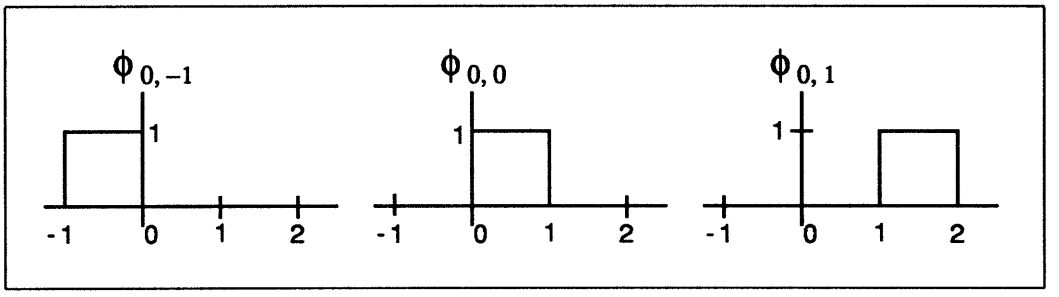

Figure 3: Orthonormal Basis for V_0 in the Haar Case

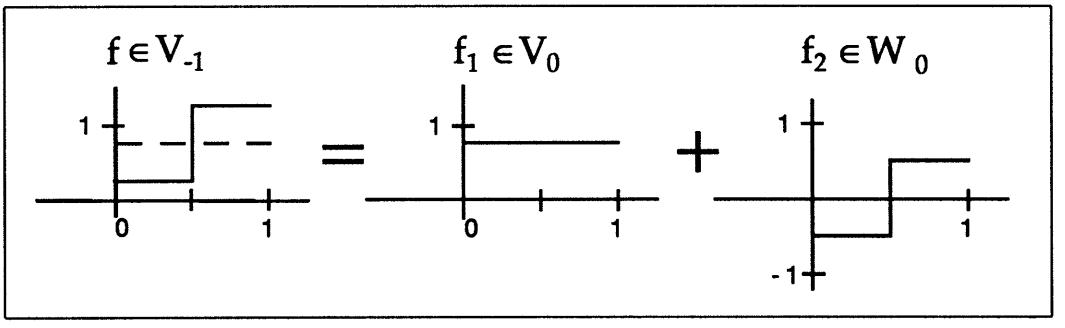

Figure 4: Decomposition of the Space V_{-1} into $V_0 \oplus W_0$

The main idea is to define, for each j, a closed subspace $W_j \subset V_{j-1}$ to be the orthogonal complement of V_j in V_{j-1}. This means that

$$V_{j-1} = V_j \oplus W_j \quad \text{and} \quad W_j \perp W_k \quad \text{if} \quad j \neq k \; .$$

It follows that $L_2 = \bigoplus_{j \in Z} W_j$ and the scaling property also carries over to the W_j's. This is shown for the Haar example in Figure 4.

The multiresolution analysis gives an orthonormal basis $\{\psi_{m,n}\}$ such that for a fixed j, $\{\psi_{j,n} \; : \; n \in Z\}$ is an orthonormal basis for W_j. The details are a bit lengthy to show but the result is the following.
Given a ϕ such that for a fixed j, $\phi_{j,n}$ is an orthonormal basis for V_j, let $h_n = \langle \phi, \phi_{-1,n} \rangle$ and $m_0(\xi) = \frac{1}{\sqrt{2}} \sum_n h_n e^{-in\xi}$. Then, a possible choice of an orthonormal basis for $L_2(R)$ is given by $\{\psi_{m,n}\}$ where

$$\hat{\psi}(\xi) = 2^{i\xi/2} \overline{m_0(\xi/2 + \pi)} \hat{\phi}(\xi/2) \rho(\xi) \tag{13}$$

and ρ is a 2π periodic function with $|\rho(\xi)| = 1$ a.e.

Thus, the multiresolution analysis starts with an orthonormal basis ϕ for V_0 and gives an orthonormal basis for L_2. The Haar example is illustrated in Figure 5.

One can, in fact, start with a (limited resolution) Riesz basis $\overline{\phi}$, use an orthogonalization trick to get an orthonormal basis ϕ (like Gramm–Schmidt for infinite dimensional spaces) and define V_0 as the span of ϕ. Many examples of this type can be found in [2]. Virtually all orthonormal bases constructed up to now can be shown to be a result of the multiresolution analysis. However, there do exist constructions of orthonormal bases which cannot be a result of multiresolution analysis. These bases have poor decay properties and it is an open question whether imposing some smoothness on $\hat{\psi}$ would eliminate such pathological cases.

6.2 Connection to Subband Coding

Writing down the equations for ψ in the time domain, one can derive an iterative algorithm for computing the DWT coefficients for a function f. In particular, it can be shown that if $\phi = \sum h_n \phi_{-1,n}$ then $\psi = \sum (-1)^n h_{-n+1}$. From this it follows that

$$\phi_{j,k} = \sum_n h_{n-2k} \phi_{j-1,n}$$

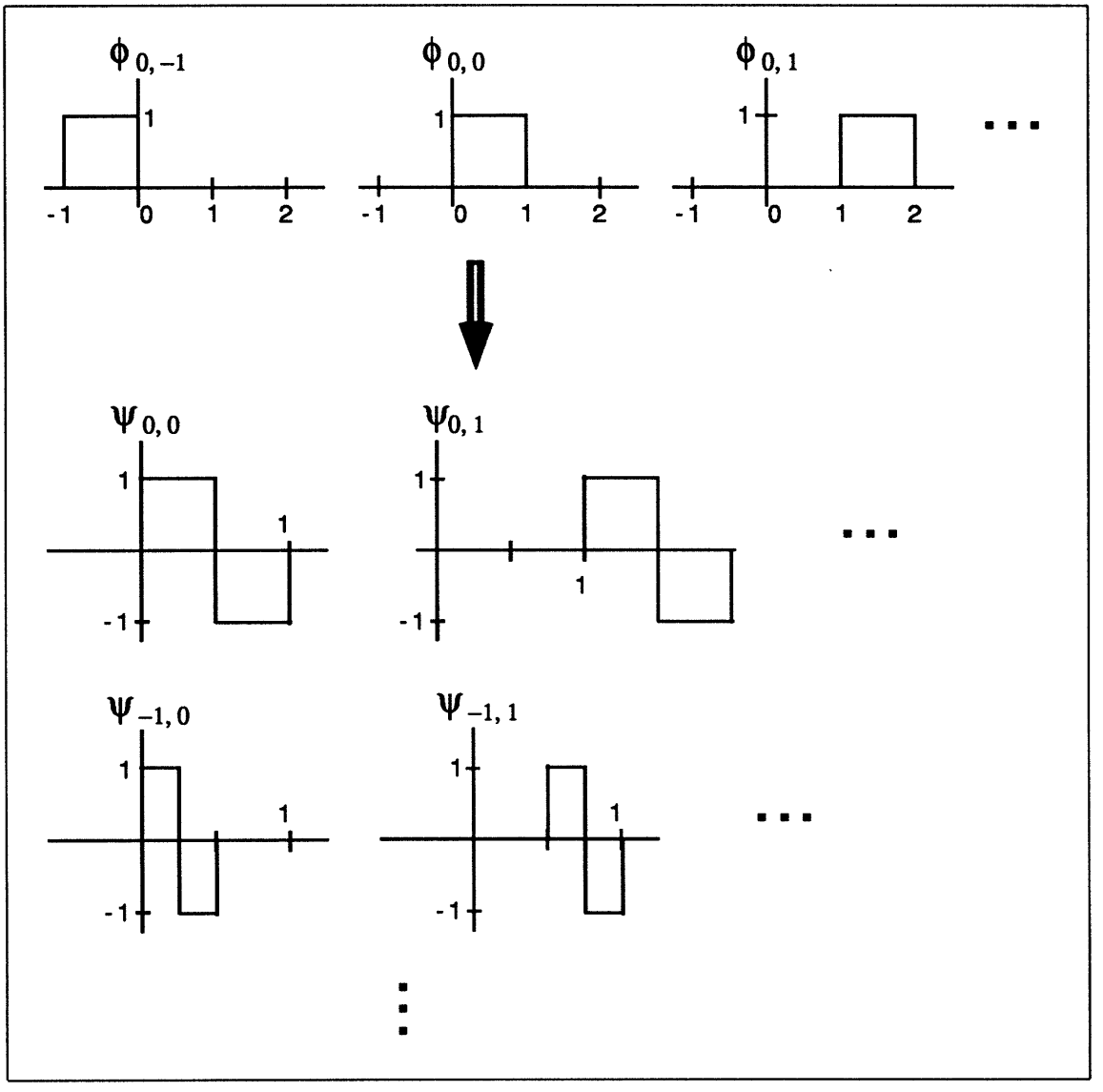

Figure 5: Haar Multiresolution Analysis

$$\psi_{j,k} = \sum_n g_{n-2k} \phi_{j-1,n}$$

where $g_n = \langle \psi, \phi_{-1,n} \rangle = (-1)^n h_{-n+1}$.
This implies that the DWT of f can be computed as follows. Starting from $\langle f, \phi_{0,n} \rangle$ one computes $\langle f, \psi_{1,n} \rangle$ and $\langle f, \phi_{1,n} \rangle$. Using the equations above, one now takes $\langle f, \phi_{1,n} \rangle$ and computes $\langle f, \psi_{2,n} \rangle$ and $\langle f, \phi_{2,n} \rangle$, etc. This can be viewed as computing coarser and coarser approximations of f along with the difference in information between adjacent resolution levels. This is equivalent to the subband coding scheme where a signal is split into its high and low frequency bands, the two signals are subsampled (decimation) and the low frequency signal is split up further, etc.

7 Compactly Supported Orthonormal Bases

The construction of orthonormal bases using multiresolution analysis leads to wavelet bases of infinite support unless the orthonormal ϕ have compact support. It can be shown that if ϕ has compact support, then m_0 must be a trigonometric polynomial. From there, one can use Bezout's theorem and spectral factorization to construct an m_0. Once a feasible m_0 is picked, $\hat{\phi}$ is given by

$$\hat{\phi}(\xi) = (2\pi)^{-1/2} \prod_{j=1}^{\infty} m_0(2^{-j}\xi)$$

This is a tight frame with frame bound of one. To insure that it is orthonormal $\|\psi\|$ must equal one and this puts another technical constraint on $m_0(\xi)$ (several equivalent conditions must hold). In general, there are no closed form solutions for ϕ's or ψ's resulting from this approach. However, Daubachies shows ways to compute values of ϕ recursively.

7.1 Related Results

Other results consider the regularity properties of various compactly supported orthonormal wavelet bases (Hölder exponent results). These can be grouped into frequency domain and time domain methods. Typically, regularity is gained at the expense of longer support. Symmetry of wavelets is also studied. One remarkable result says that the only compactly supported real orthonormal wavelet basis that is also symmetric is the Haar basis. The

way to get symmetry is by introducing the biorthogonal bases. In this case, there is the frame $\psi_{m,n}$ and the dual frame $\tilde{\psi}_{m,n}$ used for reconstruction. The additional requirements are ψ is symmetric and $\langle \psi_{m,n}, \tilde{\psi}_{k,l} \rangle = \delta_{m,k}\delta_{n,l}$. There is also a plethora of results available for biorthogonal wavelet bases. Another important topic not covered in this report is wavelet packets. This is a construction which uses multiple wavelets to cover the time–frequency plane in a manner which is best suited for the particular signals of interest.

The field of wavelets is so vast that it is impossible to even comment on all of the interesting results within the confines of this report.

8 Recent Advances in Wavelets

Recently, an entire issue of the IEEE Transactions on Signal Processing was devoted to the applications of wavelets. The research directions are broad and cover a range of topics including music, speech and image processing and detection, time–scale analysis, sampling theorems, regularity results, algorithms for fast computation of wavelet frames, more general time–scale plane tilings, adapted wavelets, and many others. The researchers include mathematicians, physicists, and engineers. Wavelet theory is being applied to everything imaginable and the field is growing at an incredible rate.

In staying with the current theme, the next section proposes a way in which wavelet theory can be applied to the field of neural networks.

9 Applications to Neural Networks

Neural networks and learning systems is a field which gained tremendous popularity in the mid 1980's and has recently gained acceptance in the "mathematically rigorous" community. There are many areas in this field and the one considered here is supervised learning (network is aware of its errors as seen by a knowledgeable observer) [7]. The viewpoint which lends itself to approximation theory is that a neural network can be thought of as a nonlinear function approximator. In this framework, the network is viewed as a nonlinear system parameterized by a vector known as the "weights". As more samples of the unknown function are disclosed to the network, the

weights are adjusted to decrease the prediction error.

Typically, there are two main issues to address. The first is choosing the architecture of the network. This will depend on the particular application and the kinds of functions which will need to be learned. The second is training the network, or deciding on a learning mechanism.

It was argued in [6] that for certain on–line applications (e.g, system identification) it is important to have a localized learning architecture. That is, each weight should affect the network output over a small subset of the input space. This means that if learning takes place over some region of input space for an extended period of time, the network does not "unlearn" the function for input values outside this region. This learning localization causes one to sacrifice generalization properties of the network. Loosely speaking, generalization is the capacity to store large amounts of information in a small number of parameters. It will be shown that the wavelet architecture along with some learning algorithm achieve a good tradeoff between the two extremes.

The network architecture typically considered in localized learning is a linear combination of spatially localized functions (e.g., radial basis functions) where the weights are comprised of the linear coefficients, and sometimes the position and spatial decay of the basis functions. When the weights are simply the linear coefficients, there exist some simple and provably convergent training algorithms such as Widrow's Least Mean Squares algorithm [8]. Such an architecture is shown in Figure 6.

The next section shows that a network consisting of multiresolution layers can be viewed as a DWT representation of the unknown function. The wavelet theory is used for choosing a basis and the learning mechanism tries to iteratively approximate the DWT coefficients. One property which is a direct consequence of using a wavelet basis is that the more "spatially global" neurons are updated more slowly than the "spatially local" ones. This is exactly what is needed for spatially localized learning behavior.

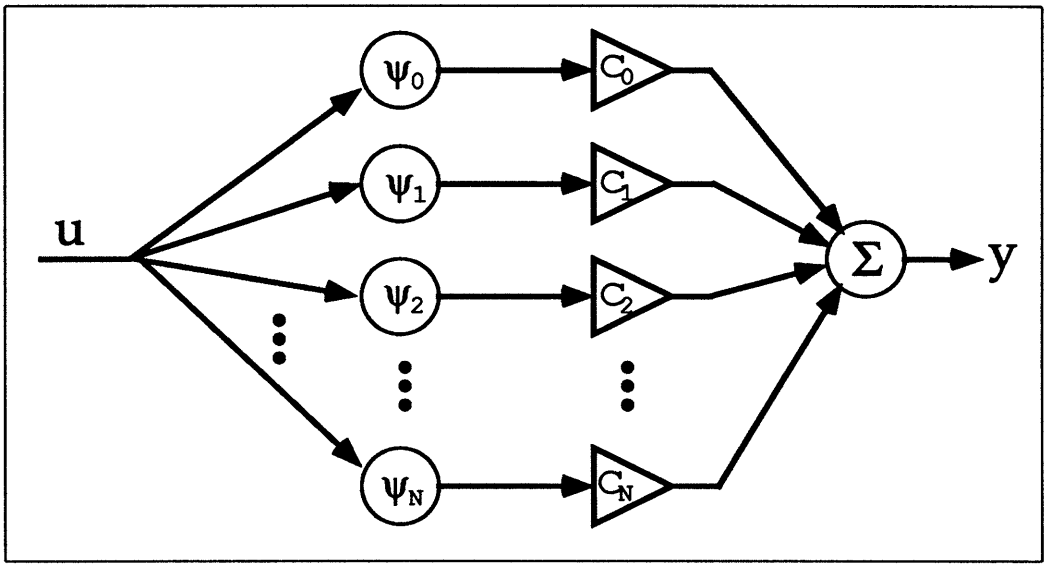

Figure 6: Neural Network Architecture

9.1 Neural Network as a Wavelet Representation

The neural network architecture given by

$$F_{\text{net}}(x) = \sum_k c_k g_k(x)$$

can equivalently be described by

$$F_{\text{net}}(x) = \sum_{m,n} c_{m,n} \psi_{m,n}(x)$$

where $\psi_{m,n}$ is the translated and dilated version of ψ as defined earlier. The spatial support of the neurons, for two dimensional inputs and $a_0 = 2$, is shown in Figure 7. Note that multiresolution does not mean multilayer, and the network output is still a linear combination of *every* neuron in the network (see Figure 6).

The results from wavelet theory characterize neuron function bases which constitute frames and possess certain regularity (if that is required of the approximation). Additionally, if one knows the spatial and frequency bounds on the function to be learned, then wavelet theory provides bounds on errors

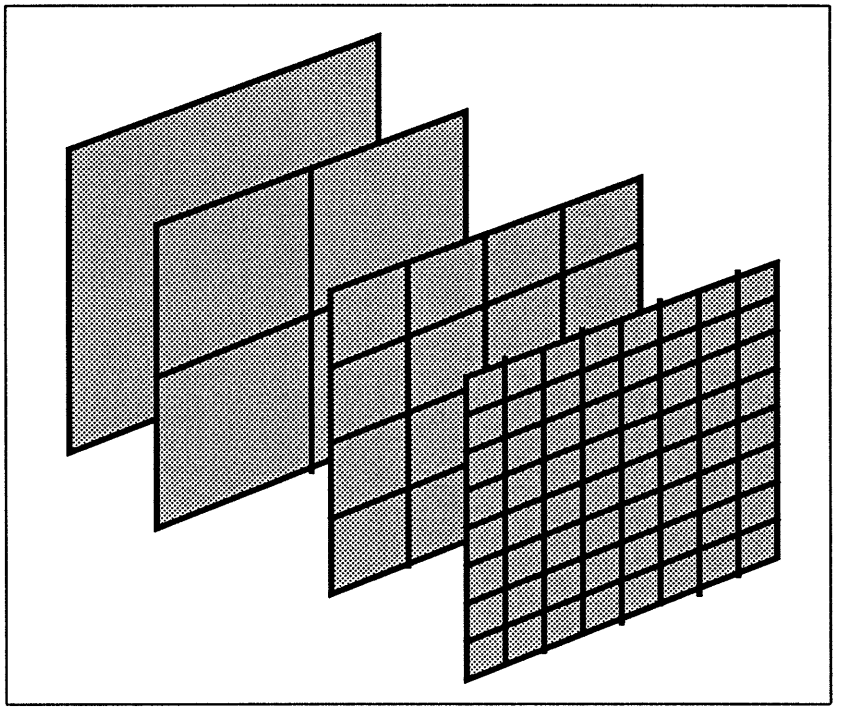

Figure 7: Spatial Multiresolution Structure

due to a truncated wavelet reconstruction. This is *extremely* important because it provides information about the size of the network given the desired accuracy. Such results are generally missing from the neural network literature, where only asymptotic (in network size), "universal approximator" results are available [9].

The remaining challenge is in learning the DWT incrementally from the samples of the unknown function f. This is tackled in the following section.

9.2 Learning Mechanisms

Once the multiresolution architecture is fixed, the neural network can be viewed as a wavelet frame representation. So far there is nothing "neural" here. The point where learning comes in is in determining the DWT coefficients. This is where the departure from the DWT begins.

In standard wavelet analysis, the function f is completely known and the DWT is computed by $\langle f, \psi_{m,n} \rangle$. In the learning case, f is not known, but as time marches on, more samples of f are disclosed to the network. The network uses this information to adjust its weights in order to decrease the prediction error. The connection to wavelets is that the optimal network weights which minimize the prediction error are precisely the DWT coefficients of f (this result was stated in Section 5.2). Hence, what is needed is a parameter update rule which will cause the weights to converge to the DWT of f.

The solution is to use an update law which is known to work for any network architecture that is linear in the parameters. The simplest of these is the Least Mean Square (LMS) algorithm.

9.2.1 LMS Algorithm

The LMS algorithm has been around for quite some time and an excellent treatment is provided in [3]. The main idea is to adjust the parameters in a way that optimally reduces the squared prediction error. This is really a gradient descent method and the parameter update law takes on the following

form:

$$c_j[k] = c_j[k-1] - \lambda \phi_j(x[k]) e[k] \tag{14}$$

where it is assumed that $F_{\text{net}}(x) = \sum c_j \phi_j(x)$, $e[k] = f(x[k]) - F_{\text{net}}(x[k])$ and λ is a small positive constant which gives the step size. Under certain persistent excitation conditions and suitable choice of λ, the LMS algorithm converges to the global optimum in this linear case. One can make this rigorous based on sufficient sampling of the input space of f and everything tending to either 0 or ∞.

The important point is that there is a simple algorithm which can incrementally update the network parameters and provide convergence to the DWT. Once the learning is complete, the functions corresponding to zero (or sufficiently small) DWT coefficients can be removed. The result is an efficient multiscale representation of the function f via a neural network.

9.3 Learning Properties

There are some nice properties associated with the wavelet architecture and the simple learning algorithm. The important property is that localized learning is maintained because parameters corresponding to the lower resolution neurons are updated much more slowly than those for the high resolution neurons. This means that when the input jumps into a new region and the prediction error becomes large, the higher resolution neurons adjust to counteract most of the error and the lower resolution neurons are adjusted more cautiously. This exactly what is needed.

On the other hand, if f is frequency band limited, the convergence of LMS guarantees that the weights of the high resolution neurons will converge to very small values even though they adapt much faster. This shows that there is also a certain amount of generalization in the network and eventually, the few lower resolution neurons will adapt if they alone are capable of approximating f.

The first property can be explained by the fact that in the LMS rule, the change in c_j is proportional to ϕ_j. But ϕ_j is proportional to $a_0^{m/2}$ where increasing m corresponds to increasing spatial resolution. In the case where

$a_0 = 2$ and the number of resolution layers used is N, for a given input x there are N neurons that are active. This means that only those N weights need to be updated. If the network "sees" an error e at this time, the error credit is assigned according to the magnitudes of these N basis functions. If λ is chosen such that the error is exactly canceled after the parameter update, the result can be viewed as the j^{th} neuron canceling $2^{j/2}/S$ of the total error e, where $S = (1 - 2^{-N/2})/(2^{1/2} - 1)$.

The second property is a simple consequence of the fact that for LMS (under mild conditions) the weights converge to the global optimum (in this case DWT).

In this way, wavelet theory along with the LMS rule provide a systematic construction of neural network architectures having certain desirable properties, and achieve an efficient representation of a function once it is learned.

10 Conclusion

This report has presented a conceptual overview of wavelets, as well as some of the important results in the field. This was followed by an application of this theory to neural networks. The report hopefully shows that the field of wavelets is vast and full of interesting results ranging from practical to the esoteric. The field is thriving because it brings together researchers from many different fields and hence, there is no shortage of innovative ideas. It is hard to imagine where wavelets will be 10 years from now, but one can be sure that fields typically dominated by Fourier techniques will undergo great change.

References

[1] Rioul and M. Vetterli,"Wavelets and Signal Processing", *IEEE Signal Processing Magazine*, Vol. 8, Oct., 1991.

[2] Ingrid Daubachies,"Ten Lectures on Wavelets", SIAM, 1992.

[3] C.R. Johnson,"Lectures on Adaptive Parameter Estimation", SIAM, 1992.

[4] K. Gröchenig,"Acceleration of the Frame Algorithm", *IEEE Trans. Signal Proc.*, Vol. 41, Dec., 1993.

[5] C. Herley, J. Kovacevic, K. Ramchandram, M. Vetterli,"Tilings of the Time–Frequency Plane: Construction of Arbitrary Orthogonal Bases and Fast Tiling Algorithms", *IEEE Trans. Signal Proc.*, Vol. 41, Dec., 1993.

[6] M. Livstone, J. Farrell and W. Baker,"A Computationally Efficient Algorithm for Training Recurrent Connectionist Networks," *Proc. Amer. Control Conf.*, Chicago, June 1992.

[7] P. Antsaklis,"Neural Networks in Control Systems", *IEEE Control Systems Magazine*, April 1992.

[8] B. Widrow and M Lehr,"30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation", *Proceedings of the IEEE*, Vol. 78, NO. 9, Sep., 1990.

[9] K. Hornik, M. Stinchcombe, and H. White, "Multilayer Feedforward Networks are Universal Approximators", *Neural Networks*, 2(5):359–366, 1989.