INSTITUTE FOR SYSTEMS RESEARCH

TECHNICAL RESEARCH REPORT

Identification of Infinite Dimensional Systems Via Adaptive Wavelet Neural Networks

by Y. Zhuang and J.S. Baras

T.R. 93-64

The Institute for Systems Research is supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), the University of Maryland, Harvard University, and Industry

Identification of Infinite Dimensional Systems Via Adaptive Wavelet Neural Networks *

Yan Zhuang[†] and John S. Baras [‡]

Institute for Systems Research and Department of Electrical Engineering
The University of Maryland, College Park, MD 20742

Abstract

We consider identification of distributed systems via adaptive wavelet neural networks (AWNNs). We take advantage of the multiresolution property of wavelet systems and the computational structure of neural networks to approximate the unknown plant successively. A systematic approach is developed in this paper to find the optimal discrete orthonormal wavelet basis with compact support for spanning the subspaces employed for system identification. We then apply backpropagation algorithm to train the network with supervision to emulate the unknown system. This work is applicable to signal representation and compression under the optimal orthonormal wavelet basis in addition to autoregressive system identification and modeling. We anticipate that this work be intuitive for practical applications in the areas of controls and signal processing.

Keywords: wavelets, neural networks, system identification, approximation.

*Research partially supported by NSF grant NSFD CDR 8803012, through the Engineering Research Center's Program and AFOSR URI grant 90-01054

[†]Email: yzhuang@src.umd.edu

[‡]Martin Marietta Chair in Systems Engineering. Email: baras@src.umd.edu

1 Introduction

There are two well known types of system identification schemes, parametric and non-parametric. The former depends on the given model structure used in identification and determines the model's parameters based on input and output of the unknown systems. The second scheme does not require the information regarding to the model structure and gives an estimate of the impulse response of the unknown systems. However, some cases are not suitable to be treated with these conventional approaches due to insufficient analytical knowledge of the plant, incomplete information on the number of key parameters and the presence of disturbance and uncertainties. Even when enough knowledge about the system is available, the model of the system may be too complicated to be used to design control systems.

We are interested in introducing another form of identification scheme which employs a parallel computational structure and uses knowledge from measurement to adapt to different models and structures. This method can be used for both linear and nonlinear system identification. The underlying idea is two-fold: first, identify the type or class of the system and pick a simple component or a structure which describes the characters of the system; second, start from the simple structure to build a basis to generate or approximate the given systems successively in an appropriate functional space.

We have found recent advancement in wavelet theory encouraging in generating an autoregressive modeling structure for system identification and signal approximation in $L^2(R)$. There have been extensive research interest and activities in wavelet theory and its applications in recent years [4] [7]. The most attractive features of wavelet theory are the multiresolution property and time and frequency localization ability. The wavelet transform decomposes a signal to its components at different resolutions. Its application actually simplifies the description of signals and provides analysis at different levels of detail. There are some successful applications of these properties in the fields of signal processing, speech processing and especially in image processing [16] [12]. It was shown [13] that it is possible to derive a base wavelet function $\psi(x) \in L^2(R)$ such that for $j, l \in Z$, $\{\psi_{j,l}(x)\}_{j,l \in Z}$ with

$$\psi_{j,l}(x) = \sqrt{2^j}\psi(2^j x - l) \tag{1}$$

is an orthonormal basis of $L^2(R)$. Any square integrable function $f(x) \in L(R^2)$ can be represented as

$$f(x) = \sum_{j,l} w_{j,l}\psi_{j,l}(x), \tag{2}$$

the coefficients $w'_{j,l}s$ carry the information of $f(x)$ near frequency 2^j and near $x = 2^{-j}l$.

Any signal in $L^2(R)$ can be decomposed to its components in different scales in subspaces of $L^2(R)$ of corresponding resolutions and the reverse is true when the regularity condition for the base wavelet $\psi(x)$ is introduced [7] [13]. The base function $\psi(x)$ plays a central role in this formulation.

We consider identification as constructing a suitable subspace of $L^2(R)$ and generating a function to approximate the output of the system with respect to the input since a large class of transfer functions of flexible structure systems and distributed systems belong to $L^2(R)$. The identification of a transfer function or a input output relation can thus be formulated as the approximation of a function in $L^2(R)$ by its projection on an appropriate subspace of $L^2(R)$. If we can construct a suitable subspace of $L^2(R)$ in an appropriate scale spanned by dilating and shifting a base wavelet function, we should be able to approximate a function in $L^2(R)$ with a function in the subspace of the relevant resolution in the sense of minimizing a norm of the difference between the two functions. Naturally, the best approximation is predetermined by the subspace in consideration and thus by the base wavelet which determines the dynamical characteristics of the subspace used for approximation. If partial information of the system is available *a priori*, or the class of the function to be approximated is detected, an appropriate wavelet basis could be built and the multiresolution property can be used to approximate the function regressively.

When a function in $L^2(R)$ or a transfer function in $H^2(R)$ is unknown, the wavelet system is feasible for its identification. Some work relating wavelets to linear systems can be found in [15]. Since a closed expression is usually not available for practical purposes, it is necessary to use a sum of finite number of functions, typically of lower order or less complexity, to approximate the original transfer function. A wavelet system can be implemented to emulate the unknown system. This process is completed by adjusting the coefficients with respect to the wavelet basis.

The transfer function of an infinite dimensional system or a distributed system is usually a sum of infinitely many functions of certain classes. Under certain conditions, a distributed system with a transfer function $G(x, \xi, s)$ can be represented by infinite many parallel aperiodic distributed blocks and oscillatory blocks[3],

$$G(x, \xi, s) = \sum_{i=1}^{\infty} G_i(s) p_i(x) q_i(\xi) \tag{3}$$

where p and q are the eigenfunctions of the corresponding boundary value problems. The Green's function has a similar structure which is the system's impulse response. We shall use $G(s)$ to denote the above transfer function for clarity in notation. When we mention a transfer function, we refer either a concrete transfer function or an implied input output relation in the rest of the paper. The summation form of both transfer functions and the Green's function can be arranged in a tree-like structure; the sum of the weighted functions can be laid out as a summation of the weighted subsums of similar structures. The weighted sum reminds us a computational structure: neural networks. We arrange the wavelet system in a similar fashion such that coefficients of the systems turn into the synapses of the neural networks. The process of approximation becomes training of the neural network. Techniques from neural networks are applicable.

A general representation of a neural network is a computational structure of finite linear combinations of the form

$$g(\mathbf{x}) = \sum_{j=1}^{N} w_j \sigma(\mathbf{a}_j^T \mathbf{x} + \mathbf{b_j}), \tag{4}$$

where $\mathbf{x}, \mathbf{a}_j \in R^N$, $b_j \in R$ are fixed. The network is formed from weighted compositions and superpositions of a single, simple nonlinear pattern or response function. The univariate function σ depends heavily on the context of the application. Neural networks have found their applications in controls and system identification. A neural network was used as an emulator and controller to control a highly nonlinear truck-trailer docking problem in [14]. Some applications of neural networks have been studied and summarized in [8] regarding modeling, identification and control structures. The nonlinear functional mapping properties of neural networks are central to their applications in system identification and controls. It has been proven [6] that a two-layer neural network can approximate a nonlinear function to an arbitrary degree of accuracy. However, the number of neurons required in the networks may far exceed the limit for practical implementations. This poses a burden for the applications in on line system identification and real time system controls. An issue in control is the dynamical nature of the system. When proper dynamics are included in the neural networks, the performance of the networks is expected to be improved. With wavelet dilations incorporated into the network, the signals to the neurons are preprocessed by the wavelet blocks. We anticipate that the information from the wavelet basis will reduce the number of neurons needed to achieve the same performance provided that the wavelets contain useful information of the systems in consideration.

Our thoughts on a unified work on wavelets and neural networks are further encouraged by the work in [20]. We are interested in a new formulation using both the multiresolution property from wavelet decomposition and the convenience of computational structures of neural networks to approximate the unknown plants; We introduce in this paper a self-tuning wavelet neural network which adjusts its wavelet basis according to measurements. We call it an adaptive wavelet neural network (AWNN).

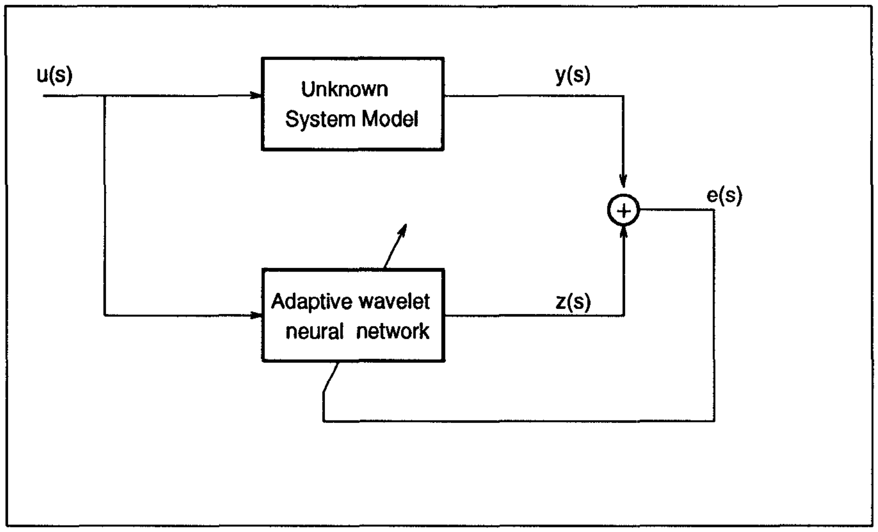

Figure 1: A wavelet neural network identification structure

This paper is organized as follows. The next section formulates the identification problem via adaptive wavelet neural networks. The third section provides details on the selection of the optimal wavelet base function for the wavelet neural networks. The fourth section addresses the network training and discusses a learning algorithm. The last section suggests future research and concludes the paper.

2 Problem statement

Given an infinite dimensional stable system with unknown transfer function $G(s)$, we set up an identification structure shown in Figure 1, in which $u(s)$ and $y(s)$ are the input and output to and from the unknown system. An adaptive wavelet neural network block (AWNN) is used to emulate the given system with $z(s)$ as its output. The matching error $e(s)$ is defined as the difference between $y(s)$ and $z(s)$. The network is tuned to match the system through minimizing the error $e(s)$.

The structure of an adaptive neural wavelet network is shown in Figure 2 , in which $u(s)$ is the input to both the system and the network, $z(s)$ is the corresponding output. This network contains a hidden layer of an appropriate wavelet basis $\{\psi_{j,l}\}$ from dilating and shifting a base wavelet $\psi(s)$ which is to be determined via an optimal adaptive scheme of basis selection. The activate function $\sigma(\cdot)$ is a nonlinear function. One of the possible forms is a sigmoidal function

$$\sigma(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}. \tag{5}$$

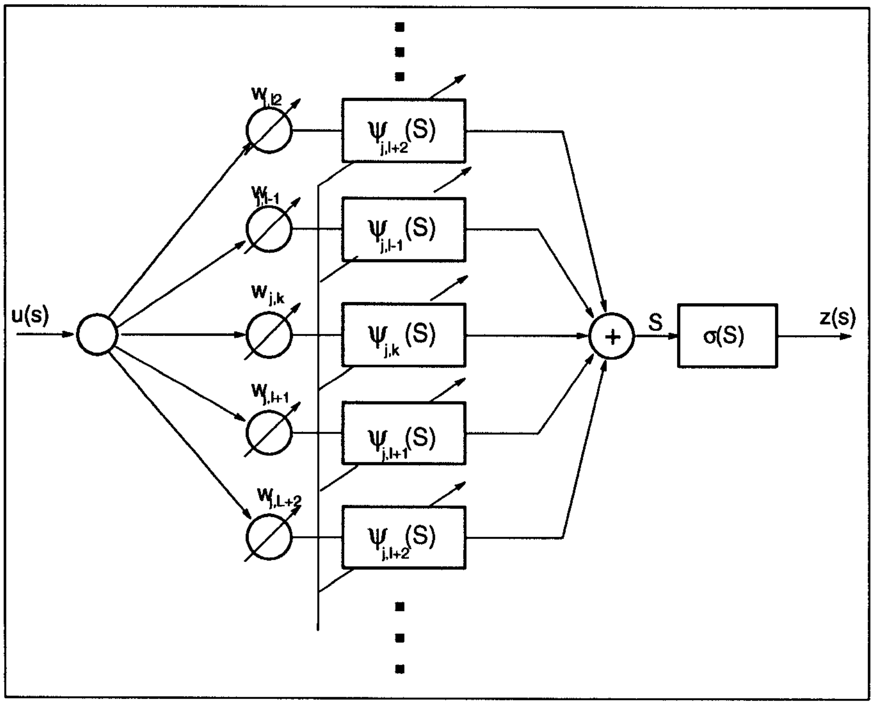

Figure 2: An AWNN block

Although this is a general setting, the dynamics of the activation function can be selected either as a linear or a nonlinear function according to the dynamics of the wavelet blocks. The output of the network is given by

$$z(s) = \sigma(\sum_{j,l} w_{j,l}\psi_{j,l}(s))u(s), \tag{6}$$

where

$$\hat{G}(s) = \sigma(\sum_{j,l} w_{j,l}\psi_{j,l}(s)) \tag{7}$$

is the estimated transfer function of the unknow system. The function $\hat{G}(s)$ approximates the transfer function to a certain level of resolution which depends on the resolution of the subspaces spanned by the wavelet functions.

The base wavelet function $\psi(x)$ determines the dynamical nature of the adaptive wavelet neural networks. How to choose the right wavelet function $\psi(x)$ is an important and non-trivial issue which has drawn recent attentions from the signal processing communities [18]. Different base wavelet functions shall generate different subspaces used to approximate transfer functions in $L^2(R)$ and produce different results. We are interested in finding the wavelet function which describes the dynamical behavior of the systems in consideration most closely. When incorporated into the network, the wavelet network should have the best performance for a certain complexity or to provide a certain performance level with a minimum complexity. This is true when $\psi(x)$ is chosen to contain information regarding the class on the given systems. We say a base wavelet is optimal if a nonnegative additive information measure $\mathcal{M}$ [5] which describes the distance between a finite length signal and the wavelet basis generated by $\psi(x)$ is minimized. The information measure is a functional which is defined by

$$\mathcal{M} : L^2(R) \times L^2(R) \longmapsto R^+. \tag{8}$$

We shall use this optimal wavelet function $\psi(x)$ in our wavelet neural networks for system approximation. The selection of the optimal base wavelet shall be discussed in detail in the section that follows.

We define the random error at instant k by the random sample (u_k, y_k) as the difference between y_k and z_k, with the system output y_k as the desired output for the neural network. The error at the k^{th} is defined by

$$e_k = y_k - z_k. \tag{9}$$

The square of error at step k is

$$E_k = \frac{1}{2}[y_k - z_k]^2. \tag{10}$$

The accumulated error E,

$$E = \sum_k E_k \tag{11}$$

sums the errors of the first k iterative steps. The network with a minimal matching error E is required to emulate the unknow system. The identification problem transforms into trajectory learning in discrete time domain.

Our problem becomes two folds: selecting the best wavelet basis for a wavelet neural network; training the AWNN afterwards to match the unknown plant. First, we need to find the optimal base wavelet function $\psi^*(x)$ such that the positive cost measure $\mathcal{M}$ is minimized for the detected dynamical behavior of a given system, i.e.,

$$\psi^*(x) = arg \min \mathcal{M}_\psi(\psi, f(x)). \tag{12}$$

Secondly, we need to train the network to emulate the given system in the sense of finding the optimal weights $\{w_{j,l}\}$ to minimize the cost index J which is

$$J_{opt} = \min_w E[w]. \tag{13}$$

The input-output relation of the trained neural wavelet network is used to represent the transfer function of the given system to facilitate the design of control systems. This forms a self-tuning system identification scheme via an AWNN.

3 Selection of base wavelet functions

We shall study the problem of choosing the optimal wavelet basis with compact support of an appropriate size in this section. We first briefly review the multiresolution property of wavelet functions and the conditions for generating a set of compactly supported discrete wavelet basis in terms of properties of quadrature mirror filter (QMF) banks [19]. We then introduce the concepts of information measure as a distance measure and the optimal discrete orthonormal wavelet basis under the information measure. A systematic approach is being developed here to derive the information gradient and the best wavelet basis. This approach can be implemented for real time systems due to our parameterization of the problem.

A multiresolution approximation due to [13] of $L^2(R)$ is a sequence $\{V_j\}_{j\in Z}$ of closed subspaces of $L^2(R)$ such that the following hold with Z denoting the set of all integers,

$$V_j \subset V_{j+1}, \ \forall j \in Z \tag{14}$$

$$\bigcup_{j=-\infty}^{+\infty} V_j \text{ is dense in } L^2(R) \text{ and } \bigcap_{j=-\infty}^{+\infty} V_j = \{0\} \tag{15}$$

$$f(x) \in V_j \Longleftrightarrow f(2x) \in V_{j+1}, \ \forall j \in Z \tag{16}$$

$$f(x) \in V_j \Longrightarrow f(x - 2^{-j}k) \in V_j, \ k \in Z \tag{17}$$

and there is a scaling function $\phi(x) \in L^2(R)$, such that, for all $j \in Z$,

$$\sqrt{2^j}\phi(2^j x - l))_{l\in Z} \tag{18}$$

is a orthonormal basis of V_j with $V_j \subset V_{j+1}$. With this setting, H_j, the complement of $V_j \subset V_{j+1}$, can be expressed as

$$V_j \oplus H_j = V_{j+1}, \tag{19}$$

with

$$V_J = \oplus_{j=-\infty}^{J-1} H_j. \tag{20}$$

For all j, there is a wavelet function $\psi(x)$, such that,

$$\sqrt{2^j}\psi(2^j x - l))_{l\in Z} \tag{21}$$

is an orthonormal basis of H_j. The additional information in an approximation at resolution 2^{j+1} compared with the resolution 2^j is contained in the subspace H_j, the orthogonal complement of $V_j \in V_{j+1}$. If we define P_{V_j} to be a projection operator in $L^2(R)$ and I to be the identity operator, then

$$P_{V_j} \rightarrow I, \text{ as } j \rightarrow +\infty. \tag{22}$$

A particular useful setup for our problem is a set of discrete orthonormal wavelet basis with compact support. It is useful for real time implementation on digital computers. The compactness of support provides a means of isolation and detection of signals at a certain region which has proven useful in signal processing communities. Both the discrete scaling function $\phi(x)$ and the discrete wavelet function ψ can be parameterized by a set $\{c_k\}$ with ks belonging to a set of integers.

The scaling function $\phi(t)$, with t denoting discrete time, compactly supported on $[0, K-1]$, can be expressed as [7]

$$\phi(t) = \sum_k c_k \phi(2t - k). \tag{23}$$

The discrete wavelet is given by

$$\psi(t) = \sum_k d_k \phi(2t - k), \tag{24}$$

where

$$c_k \neq 0, \; k \in [0, K-1]. \tag{25}$$

These are the two fundamental equations for wavelet function $\psi(t)$. The scaling function $\phi(t)$ can be nonzero only on $[0, K-1]$ due to the finite duration of the sequence $\{c_k\}$. The base wavelet function obtained through $\phi(t)$ is also compactly supported. The coefficients $\{c_k\}$ and $\{d_k\}$ can be identified as a low pass filter and a high pass filter respectively. Let us denote $h_0(k) = c_k/2$ and $h_1(k) = d_k/2$ and take their Fourier transforms

$$H_0(e^{j\omega}) = \sum_k h_0(k) e^{-j\omega}, \tag{26}$$

and

$$H_1(e^{j\omega}) = \sum_k h_1(k) e^{-j\omega}. \tag{27}$$

The conditions for compactly supported orthonormal wavelet and scaling functions are equivalent to that the matrix

$$\mathcal{H}(\omega) = \begin{bmatrix} H_0(e^{j\omega}) & H_1(e^{j\omega}) \\ H_0(e^{j(\omega+\pi)}) & H_1(e^{j(\omega+\pi)}) \end{bmatrix} \tag{28}$$

is unitary for all ω for the two-channel quadrature mirror filter (QMF) bank [1]. This is the constraint that the parameters c_k should satisfy. In particular, the cross-filter orthonormality implied by the unitary property, is satisfied by the choice [1]

$$H_1(z) = z^{K-1} H_0(-z^{-1}), \text{ N even} \tag{29}$$

or in the time domain,

$$h_1(k) = (-1)^{k+1} h_0(K-1-k), \tag{30}$$

and in addition

$$\int \psi(t)dt = 0. \tag{31}$$

As we can see from the above, both the scaling function and the wavelet function depend on the choice of $\{c_k\}$ for $k \in [0, K-1]$. The base wavelet function depends on the selection of this set of parameters.

The key to choosing the optimal wavelet base for the AWNN lies in the appropriate parameterization and the right performance measure in addition to the accurate interpretation of physical phenomena. A method is proposed in [18] [10] for choosing a wavelet for signal representation based on minimizing an upper bound of the L^2 norm of error in approximating the signal up to the desired scale. Coifman et al. derived an entropy based algorithm for selecting the best basis from a library of wavelet packets [5]. However, a direct method to systematically generate the best orthonormal discrete wavelet basis with compact support is still to be developed. We shall provide here a direct approach to calculate the best discrete wavelet basis.

We first introduce a distance measure for optimization purpose. Inspired by the work in [5], we define an additive information measure of entropy type and the optimal basis as the following.

Definition 3.1 *A non negative map $\mathcal{M}$ from a sequence $\{f_i\}$ to R is called an additive information measure if $\mathcal{M}(0) = 0$ and $\mathcal{M}(\sum_i f_i) = \sum_i \mathcal{M}(f_i)$.*

Definition 3.2 *Let $x \in R^N$ be a fixed vector and $\mathcal{B}$ denote the collection of all orthonormal bases of dimension N, a basis $B \in \mathcal{B}$ is said to be optimal if $\mathcal{M}(Bx)$ is minimal for all bases in $\mathcal{B}$ with respect to the vector f.*

We shall define a distance measure between a signal and its decompositions to subspaces of $L^2(R)$ motivated by Shannon entropy (Shannon's formula) [9]

$$H(X) = H(P) = -\sum_{x \in X} P(x) \log P(x), \tag{32}$$

which is interpreted as a measure of the information content of a random variable X with distribution $P_x = P$ in information theory.

Definition 3.3 *Let H be a Hilbert space which is an orthogonal direct sum*

$$H = \oplus \sum H_i, \tag{33}$$

a map $\mathcal{E}$ is called decomposition entropy if

$$\mathcal{E}(v, \{H_i\}) = -\sum \frac{\|v_i\|^2}{\|v\|^2} \log \frac{\|v_i\|^2}{\|v\|^2} \tag{34}$$

for $v \in H$, $\|v\| \neq 0$, such that

$$v = \oplus \sum v_i, v_i \in H_i, \tag{35}$$

and we set

$$p \log p = 0, \ \textit{when } p = 0. \tag{36}$$

Entropy is a good measure for signal concentration in signal precessing and information theory. The value of $\exp \mathcal{E}(v)$ is proportional to the number of coefficients and the length of code words necessary to represent the signal to a fixed mean error and to error less coding respectively. The number $\frac{\|v_i\|^2}{\|v\|^2}$ is the equivalent probability measure in the decomposition entropy. In our system identification formulation, energy concentration is identified with model of lower order or networks with less complexity.

Let $\psi(t)$ be the base wavelet function and let $\Psi(t)$ represent the orthonormal discrete wavelet basis of L^2 generated by dilation and shifting of $\psi(t)$, similarly, we define Ψ_j to be the basis of H_j. We write $\Psi(t) = \{\psi_{j,l}(t)\}$ and $\Psi_j(t) = \{\psi_{j,l}(t)\}_{l \in Z}$ respectively. We treat both $\Psi(t)$ and Ψ_j as operators and thus define the following.

Definition 3.4 *Let Ψ be a basis given above, a base operation is defined to be a map from $L^2(R)$ to a set of real numbers, i.e., $\Psi(t)f(t) = \{f_{j,l}\}_{j,l \in Z}$ where $f_{j,l} = \langle f(t), \psi_{j,l}(t) \rangle$ for all $f(t) \in L^2$.*

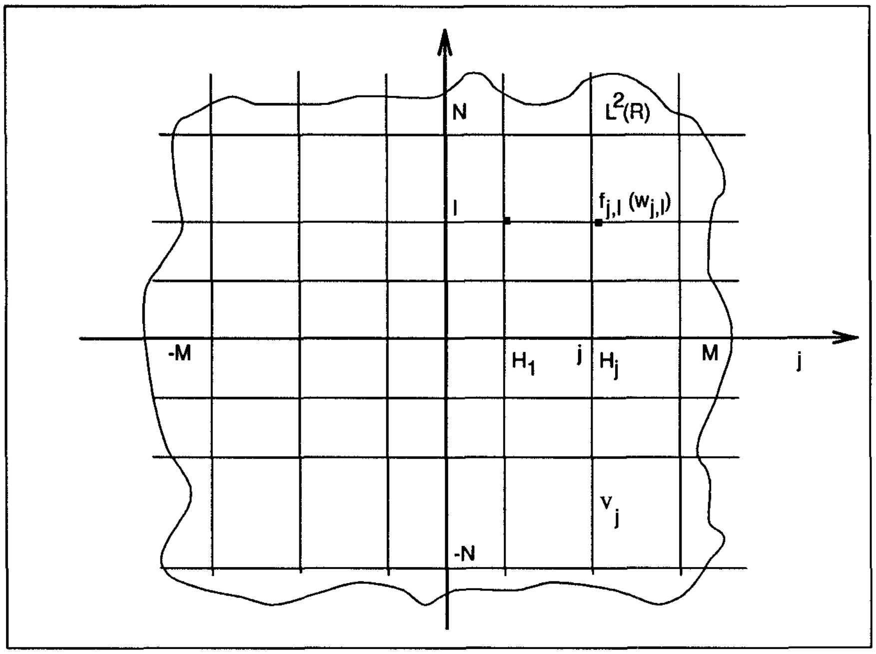

Figure 3: Mesh structure of the projection space

Consider V_J, the subspace of $L^2(R)$, with

$$V_J = \oplus_{j=-\infty}^{J-1} H_j, \tag{37}$$

Equation (7) and Equation (2), let M and N be appropriate positive integers, we truncate the approximation in Equation (2) to a scale up to M, we have

$$f(x) = \sum_{j=-M}^{M} \sum_{l=-N}^{N} w_{j,l} \psi_{j,l}(x). \tag{38}$$

The subspaces used to approximate function $f(x)$ has a mesh of size $(2M+1) \times (2N+1)$ as in Figure 3.

Given a function or signal $f(t) \in L^2(R)$ and a base wavelet function $\psi(t)$ with a finite mesh of size $(2M+1) \times (2N+1)$, we can decompose the signal to the orthogonal subspaces as

$$f(t) = \sum_{j=-M}^{M} \sum_{l=-N}^{N} f_{j,l} \psi_{j,l}(t). \tag{39}$$

We are going to find the best wavelet base function $\psi(t)$ for a given signal $f(t)$ such that the additive information measure $\mathcal{M}$ is minimized. The result of the base operation $\Psi f(t)$ appears as the weights on the nodes of the mesh. The weights on the vertical line with coordinate j is the number set produced by $\psi_j f(t)$

Although the decomposition entropy is a good measure for the "distance", it is not an additive type of map because of the norm $\|v\|$ is used to scale the vector. We thus further introduce a functional

$$\lambda(\Psi v) = -\sum_j \|v_i\|^2 \log \|v_i\|^2, \tag{40}$$

which relates to the decomposition entropy through

$$\mathcal{E}(v, \{H_i\}) = \|v\|^{-2} \lambda(\Phi v) + \sum_j \log \|v\|^2. \tag{41}$$

The former function in (40) is an additive measure. Since minimizing the later minimizes the later, we minimize functional $\lambda(\Phi f)$ for seeking the optimal wavelet basis through multiresolution decompositions.

The weight of decomposition of signal $f(t)$ on a subspace H_j is measure by a subnorm $\|f_j\|$ defined as

$$\|f_j(t)\| = \left\| P_{H_j}[f(t)] \right\|, \tag{42}$$

where

$$\|f_j\|^2 = \sum_{l=-N}^{N} f_{j,l}^2. \tag{43}$$

Similarly, the norm of the decomposed signal is given by

$$\|f(t)\|^2 = \sum_{j=-M}^{M} \|f_j\|^2. \tag{44}$$

We need to further find $\frac{\partial f_{j,l}}{\partial c_k}$ which is a measure of sensitivity of the component of signal decomposition to a wavelet basis versus the change of the defining parameter set of the base wavelet. One can solve this through numerical methods from the relations and definitions. Based on the definition of information gradient and the properties of QMF discussed earlier, we derive an explicit expression as following.

Lemma 3.1 *The sensitivity gradient* $\frac{\partial \psi_{j,l}}{\partial c_k}$ *of component* $\psi_{j,l}$ *of the wavelet basis* Ψ *versus parameter* c_k *is given by*

$$\frac{\partial \psi_{j,l}}{\partial c_k} = \sqrt{2^j} \sum_n \left[(-1)^{K-k} \phi(2^{j+1}t - 2l - n) + (-1)^{n+1} c_{K-1-n} \frac{\partial}{\partial c_k} \phi(2^{j+1}t - 2l - n) \right]. \tag{45}$$

Proof:
From the fundamental equation of wavelets (24),

$$\frac{\partial \psi_{j,l}}{\partial c_k} = 2\frac{\partial}{\partial c_k}\sqrt{2^j}\sum_n h_1(n)\phi(2^{j+1}t - 2l - n). \tag{46}$$

This is

$$\frac{\partial \psi_{j,l}}{\partial c_k} = 2\sqrt{2^j}\sum_n \left[\frac{\partial h_1(n)}{\partial c_k}\phi(2^{j+1}t - 2l - n) + h_1(n)\frac{\partial}{\partial c_k}\phi(2^{j+1}t - 2l - n)\right]. \tag{47}$$

From Equation (23), we have

$$\frac{\partial \phi(t)}{\partial c_k} = \phi(2t - k). \tag{48}$$

hence,

$$\frac{\partial}{\partial c_k}\phi(2^{j+1}t - 2l - n) = \phi(2^{j+2}t - 4l - 2n - k). \tag{49}$$

We need to find $\frac{\partial h_1(n)}{\partial c_k}$, from the time domain relation (30) of the QMF, we have ,

$$h_1(n) = (-1)^{n+1}h_0(K - 1 - n) \tag{50}$$

with h_0 being compactly supported on $[0, K-1]$. Thus,

$$h_1(n) = \frac{1}{2}\frac{\partial}{\partial c_k}(-1)^{n+1}c_{K-1-n}, \tag{51}$$

there is only one nonzero term when $K - 1 - n = k$. This yields,

$$\frac{\partial h_1(n)}{\partial c_k} = \frac{(-1)^{K-k}}{2}. \tag{52}$$

The lemma is proven through (49) and (52).

□

This lemma establishes a direct link between the rate of change of the components in the basis Ψ and the variations of the parameters in the fundamental equations of wavelets, which leds to the next theorem. We introduce the following theorem to show the relationship between the information measure and the parameter set c_k and the relation here shall provide a clue for developing an algorithm to find the optimal base wavelet function for the AWNN.

Theorem 3.1 *Let $\lambda(\cdot)$ be the additive information measure and $[0, K-1]$ be the compact support for $\{c_k\}$ and Ψ be the corresponding wavelet basis, let $f(t)$ be a fixed signal in $L^2(R)$, then the gradient of the information measure with respect to the set $\{c_k\}$ for the signal is given by*

$$\begin{aligned}\frac{\partial \lambda(\Psi f(t))}{\partial c_k} &= -2\sum_j \sum_l \log 2 \|f_j\|^2 \cdot \sqrt{2^j} \sum_n \Big[(-1)^{K-k} \Big\langle f(t), \phi(2^{j+1}t - 2l - n) \Big\rangle \\ &\quad + (-1)^{n+1} c_{K-1-n} \Big\langle f(t), \phi(2^{j+2}t - 4l - 2n - k) \Big\rangle \Big] . \end{aligned} \tag{53}$$

Proof:
By the chain rule, we have the information gradient

$$\frac{\partial \lambda(\Psi f(t))}{\partial c_k} = \sum_j \frac{\partial \lambda(\Psi f(t))}{\partial \|f_j\|^2} \frac{\partial \|f_j\|^2}{\partial c_k}. \tag{54}$$

The definition of information measure $\lambda(f(t))$ in (40) yields,

$$\begin{aligned}\frac{\partial \lambda(\Psi f(t))}{\partial \|f_j\|^2} &= -\log \|f_j\|^2 - 1 \\ &= -\log 2 \|f_j\|^2 , \end{aligned} \tag{55}$$

with 2 being the base of *log* function. We use the chain rule again,

$$\begin{aligned}\frac{\partial \|f_j\|^2}{\partial c_k} &= \frac{\partial}{\partial c_k} \sum_l f_{j,l}^2 \\ &= 2 \sum_l f_{j,l} \frac{\partial f_{j,l}}{\partial c_k}. \end{aligned} \tag{56}$$

We have so far

$$\frac{\partial \lambda(\Psi f(t))}{\partial c_k} = -2 \sum_j \sum_l \log 2 \|f_j\|^2 f_{j,l} \frac{\partial f_{j,l}}{\partial c_k}. \tag{57}$$

Since

$$\frac{\partial f_{j,l}}{\partial c_k} = \left\langle f(t), \frac{\partial \psi_{j,l}}{\partial c_k} \right\rangle , \tag{58}$$

the result from the previous lemma concludes the proof.

□

The theorem demonstrates an explicit relation among the gradient of the additive information measure, parameter set $\{c_k\}$ and the measured signal $f(t)$. It will facilitate the search for the optimal wavelet basis due to our parameterization and the information measure.

We have identified the problem of finding the optimal wavelet basis Ψ with that of finding a parameter set $\{c_k\}$ such that the additive information measure λ is minimized. Once the set $\{c_k\}$ is determined, both the scaling function ϕ and the base wavelet function ψ can be derived afterwards. Equipped with the above theorem, the information gradient is available, different optimization schemes can be applied to solve this problem. We have developed a basis choosing algorithm based on a steepest descent method as follows. To simplify notation, we denote the parameter set $\{c_0 c_1 \cdots c_{K-l}\}$ by a vector C.

Algorithm 1 *Computation of the optimal wavelet basis*

Step 1: *Set* $i := 1$,
$\lambda_0 := 0$,
mesh parameters M, N;
Initialize vector C_0;
Input $f(t)$.
Step 2: *If* C_i *dose not satisfy the constraint,*
then, modify C_i *and repeat Step 2.*
Step 3: $C_i := C_{i-1} + p_{i-1} \frac{\partial \lambda}{\partial C_{i-1}}$.
Step 4: *Compute* ϕ *and* ψ.
Step 5: *Compute* λ.
Step 6: *If* $|\lambda_i - \lambda_{i-1}| > \epsilon$,
$i := i + 1$, *go to Step 2.*
Step 7: *Output the optimal basis* Ψ *and stop.*

The mesh size is governed by the choice of parameter M and N. Obviously, when M and N turn to infinity, the supporting subspace spanned by the dilations and shifts of the base wavelet turns to space $L^2(R)$. The size of the mesh is identified with the complexity of the resulted AWNN. The constraint on the parameter c_k is dominated by the unitary property of the QMF bank which can be transformed into an algebraic equation.

This section has provided us a direct method to construct an optimal orthonormal wavelet basis with compact support. The parameterization of both the information measure and the base wavelet allows an explicit expression of information gradient with respect to the optimization parameters and thus paves the way to the efficient basis choosing algorithm. This methodology of the optimal basis selection in a general setting is useful not only within this identification structure but also to signal approximation and reconstruction in $L^2(R)$. The parametrization of cost functionals is not unique, other forms of measures or cost functions may be introduced according to the contexts of the actual physical problems.

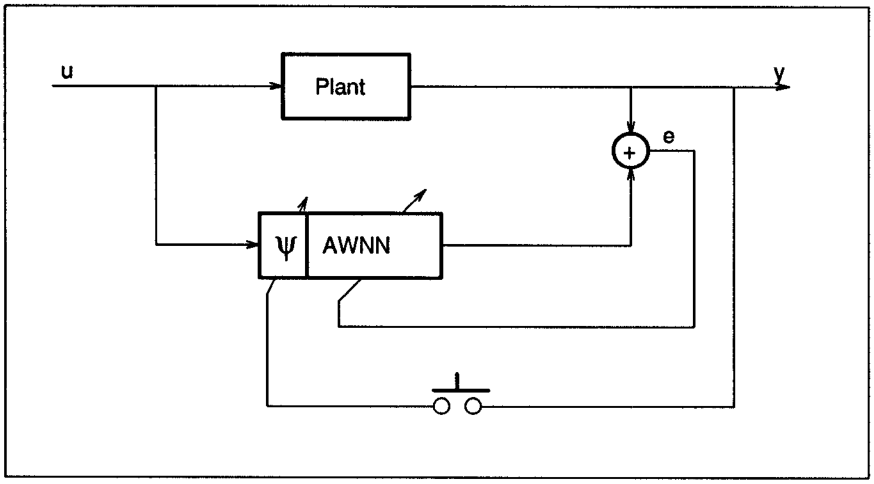

Figure 4: AWNN training structure

4 Network training

This section describes a supervised learning process of the AWNN. The training of an AWNN consists two stages, a pre-training procedure and an actual training scheme via weight updating. The pre-training is a preparation process of configuration or adjusting the basis of the network based upon the output measurements from the unknown systems excited by a test signal. The purpose is to equip the network with the appropriate dynamics and generate an AWNN of a manageable size. The network is trained afterwards with a supervised learning process. The training structure is shown in Figure 4.

During the first stage, the network takes the output of the unknow systems excited by a test signal and looks for the best wavelet basis with the switch at the closed position. The algorithm given in the previous section is used here to generate the best wavelet basis Ψ for the AWNN. The dynamical behavior of the AWNN is thus determined by this process. This stage also provides appropriate initial weights for the network training to start with. Since the basis contains the measured information of the unknown system, the required size of the network is reduced compared with a neural network without the dynamical components. This will speed up the network training process.

The next stage is the network training which is a goal-directed learning aimed at minimizing the relevant cost functional. It is supervised learning since certain pattern of $\psi(x)$ related to the unknown system is used during training. Different training algorithms were discussed in [11], [17] and [2]. Due to the convenience of our problem formulation, we use the backpropagation algorithm in [11] to train the AWNN. The backpropagation algorithm, an extension of LMS algorithm, modifies the weights at each step with nonlocal error information. This is an implied feedback which closes the loop for adapting weights of the AWNN. The backpropagation provides a suboptimal solution in the sense of using a finite number of wavelet blocks to approximate the infinite dimensional system. The task here is to minimize the cost functional J of Equation (13) which is rewritten here for convenience.

$$J_{opt} = \min_{w} E[w]. \tag{59}$$

From the structure of the AWNN, we have

$$S = \sum_{j,l} w_{j,l}\psi_{j,l}(s)u(s) \tag{60}$$

as the input to the sigmoidal function $\sigma(\cdot)$. We update the weight $w_{j,l}(k)$ at k^{th} iteration by a stochastic difference equation

$$w_{j,l}(k+1) = w_{j,l}(k) + q_k \Delta w_{j,l}(k) \tag{61}$$

where

$$\Delta w_{j,l}(k) = -\frac{\partial E_k}{\partial w_{j,l}}. \tag{62}$$

with the learning coefficients q_k's satisfying,

$$\sum_k q_k = \infty \tag{63}$$

$$\sum_k q_k^2 < \infty. \tag{64}$$

The condition (63) constrains the sequence $\{q_k\}$ to decrease slowly, while (64) constrains to decrease q_k quickly. The combined effect is to guarantee the appropriate learning rate.

The gradient of the cost functional with respect to the weight $w_{j,l}$ is expressed as

$$\frac{\partial J}{\partial w_{j,l}} = \sum_k \frac{\partial E_k}{\partial w_{j,l}}. \tag{65}$$

We refer the definition of the square of error at step k in Equation (10) and use the subscript k of a variable to denote the value of the variable at the instant k. By the chain rule, we have

$$\begin{aligned} \frac{\partial E_k}{\partial w_{j,l}} &= -(y_k - z_k)\frac{\partial z_k}{\partial w_{j,l}} \\ &= -(y_k - z_k)\frac{\partial z_k}{\partial S_k}\frac{\partial S_k}{\partial w_{j,l}} \\ &= -(y_k - z_k)\sigma'(S_k)\psi_{j,l}u_k. \end{aligned} \tag{66}$$

Hence

$$\begin{aligned}\Delta w_{j,l} &= (y_k - \sigma(S_k))\sigma'(S_k)\psi_{j,l}u_k \\ &= (y_k - \sigma(\sum_{j,l} w_{j,l}\psi_{j,l}u_k))\sigma'(\sum_{j,l} w_{j,l}\psi_{j,l}u_k)\psi_{j,l}u_k \end{aligned} \tag{67}$$

as the weights updating scheme. The general backpropagation algorithms can be found in [11]. This process starts by assigning $y_{j,l}$, the coefficients from the base operation Ψy of the measured output $y(s)$ to the wavelet basis of the AWNN, to $w_{j,l}(0)$. The trained neural wavelet network shall be used to implement control system design. The reconstruction form the given wavelet base is the approximation of the plant up to a certain resolution. Summerizing the above yields the following algorithm.

Algorithm 2 *AWNN training scheme*

Step 1: *Set* $i := 1$,
$J_0 := 0$,
Input Ψ.
Set $w_{j,l}(0) := y_{j,l}$;
Step 2: $w_{j,l}(i) := w_{j,l}(i-1) + q_{i-1}\Delta w_{j,l}(i-1)$.
Step 3: *Compute* J_i.
Step 4: *If* $|J_i - J_{i-1}| > \epsilon$,
$i := i+1$, *go to Step 2.*
Step 7: *Stop.*

Neural networks are just another way of curve fitting to available data. They have both advantages and disadvantages. They are conceptually simple and easy to use and are adaptable to complicated problems or suitable to deal with problems which do not have a modeled structure or are too complicated to model. Another advantage is that neural networks offer a distributed, parallel processing ability thus provide integrity and possible fault tolerance. The function of each neuron is usually a simple function which is easy to implement. The most obvious disadvantage is that neural networks do not recognize and preserve the structures of the systems they deal with and there is no systematical way to determine the structures of the networks. Embedding dynamical components depending on the problem context into the networks will be useful in overcoming the disadvantages. Our attempt in designing an AWNN will be of research potential in this regard.

The AWNN can be structured differently. For example, instead of using only one hidden layer, we can use a multi-layer neural network. One of the structures is a two layer format with each neurons in the hidden layer being responsible for a subspace of fixed scale while the neuron in output layer summing the results from all the subspaces. This structure may facilitate the computation. Different computational structures are to be compared for the best result.

5 Conclusions

We have developed an algorithm for identification of infinite dimensional systems via an adaptive wavelet neural network. We first solve the problem of selecting the compactly supported optimal wavelet base function for spanning the subspaces in which the unknown system is approximated up to a predetermined resolution. An algorithm is given for constructing the optimal basis Ψ for the network emulator based on the measurements of the output from the unknown system. We then apply a backpropagation algorithm to train the resulting AWNN for system approximation. This is an efficient way of approximating an infinite dimensional system up to a certain resolution in a subspace of $L^2(R)$ spanned by the dilations and shifts of the optimal base wavelet. Our method combines the advantage of multiresolution property of wavelet decompositions and the convenience of the computational structures of neural networks. The marriage of the best from both fields should provide a powerful took kit for solving problems of a much wider range. Our approach can be generated to N dimensional case with signals from $L^2(R^N)$. The methodology developed in this paper is expected to be useful not only for system identification and autoregressive modeling but also for signal classification, signal compression and reconstruction as well. Future research is needed on these aspects.

References

[1] A.N. Akansu and R.A. Haddad. *Multiresolution Signal Decomposition.* Academic Press, Inc., 1992.

[2] S. Billings and S. Chen. Neural networks and system identification. In K. Warwick, G.W. Irwin, and K.J. Hunt, editors, *Neural Networks for Control and Systems*, chapter 9, pages 181–205. Peter Peregrinus Ltd., London, United Kindom, 1992.

[3] A.G. Butkovskiy. *Structural Theory of Distributed Systems.* Ellis Horwood Ltd., 1983.

[4] C.K. Chui. *Wavelets: A tutorial in Theory and Applications.* Acamemic Press, Inc., 250 Sixty Ave, San Diego, CA 92101, 1992.

[5] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. on Information Theory*, 38(2):713–718, Mar. 1992.

[6] G. Cybenko. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

[7] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, 3600 University City Science Center, Philadelphia PA, 1992.

[8] K.J. Hunt and D. Sbarbaro. Studies in neural network based control. In K. Warwick, G.W. Irwin, and K.J. Hunt, editors, *Neural Networks for Control and Systems*, chapter 6, pages 95–122. Peter Peregrinus Ltd., London, United Kindom, 1992.

[9] I. Csiszár and J. Körner. *Information Theory.* Akadémiai Kiadó, Budapest, Hungary, 1981.

[10] P. Jorgensen. Choosing discrete orthogonal wavelets for signal analysis and approximation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages III–308–311, Minnesota, Minneapolis, April 27-30 1993.

[11] B. Kosko. *Neural Networks and Fuzzy Systems.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1992.

[12] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. on Pattern Analysis and Machine intelligence*, 14(7):710–732, July 1992.

[13] S.G. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Transactions of The American Mathematical Society*, 315(1):69–87, Sept. 1989.

[14] D.H. Nguyen and B. Widrow. Neural networks for self-learning control systems. *IEEE Control Systems Magazine*, 10(3):18–23, April 1990.

[15] Y.C. Pati. *Wavelets and Time-Frequency Methods in Linear Systems and Neural Networks.* PhD thesis, University of Maryland, College Park, MD 20742, 1992.

[16] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, pages 14–38, Oct. 1991.

[17] N. Tepedelenlioglu, A. Rezgui, R. Scalero, and R. Rosario. Fast algorithms for training multilayer perceptrons. In B. Souček, editor, *Neural and Intelligent Systems Integration*, chapter 4, pages 107–133. John Wiley & Sons. Inc., 1991. Sixth-Generation Computer Technology Series.

[18] A.H. Tewfik, D. Sinha, and P. Jorgensen. On the optimal choice of a wavelet for signal representation. *IEEE Trans. on Information Theory*, 38(2):747–765, Mar. 1992.

[19] P.P. Vaidyanathan. *Multirate Systems and Filterbanks.* Prentice hall Signal Processing Series. P T R Prentice-Hall, Inc., Englewood Cliffs, NJ 07632, 1993.

[20] Q. Zhang and A. Benveniste. Wavelet networks. *IEEE Trans. on Neural Networks*, 3(6):889–898, Nov. 1992.