SCALAR WAVELET NETWORKS FOR TIME SERIES PREDICTION

Bruce A. Whitehead
University of Tennessee Space Institute
Tullahoma, TN 37388

Abstract

$1\frac{1}{2}$ - layer neural networks whose basis functions are scalar wavelets, trained with the quickprop gradient descent algorithm, appear to train more quickly and yield lower prediction errors on the Mackey-Glass chaotic time series than comparable networks of conventional sigmoid units.

1. INTRODUCTION

A "$1\frac{1}{2}$ - layer" neural network attempts to approximate an arbitrary real-valued multivariate function f as a linear combination of basis functions $h_i, i = 1, \ldots, m$. The basis functions h_i are typically constrained to be affine transformations of a prototype basis function h. Wavelet basis functions, by virtue of being localized in both the input-space domain and its corresponding frequency domain, form theoretically attractive candidates for the prototype basis function h. Regularly-spaced grids of wavelets, arranged in a hierarchy of meshes at successively finer resolutions, have been shown capable of approximating a large class of target functions f to any desired accuracy (Daubechies, 1990). Such a hierarchical grid, termed a *frame*, may be implemented by a feedforward neural network based on either wavelet (Bakshi & Stephanopoulos, 1992) or sigmoid (Pati & Krishnaprasad, 1993) activation functions, and the resolution of the frame can be adaptively varied in different regions of the input space (Pati & Krishnaprasad, 1993).

Regularly-spaced frames of basis functions are subject to the "curse of dimensionality," however (Pati & Krishnaprasad, 1993). Frames of wavelets have been used extensively to approximate functions of one or two variables (Meyer, 1992), but as the number of variables increases, the required number of basis functions grows exponentially.

The remainder of this paper considers adaptively-placed wavelet basis functions as a potential solution to this problem. Section 2 reviews neural networks based on wavelets and explains why scalar wavelets might be considered. Section 3 gives the procedure followed to evaluate the benchmark performance of such networks, and Section 4 discusses the results obtained.

2. WAVELET NETWORKS

A neural network approach to overcoming exponential growth in the number of basis functions required to cover n-dimensional spaces would be to replace the regularly-spaced wavelet frame with an adaptively-placed set of basis functions. The positions, scales, and orientations of these basis functions could be adjusted iteratively using an appropriate gradient-descent learning rule. The additional degrees of freedom provided by adaptable basis functions could conceivably lessen the number of basis functions required to achieve a given approximation accuracy.

A theoretical framework has been constructed for such an adaptive *wavelet network* (Zhang & Benveniste, 1992) which approximates a target function f as a linear combination of shifted scaled rotations of a prototype wavelet ψ:

$$f(\mathbf{x}) \approx g(\mathbf{x}) = w_0 + \sum_{i=1}^{m} w_i \psi(D_i R_i(\mathbf{x} - \mathbf{t}_i)) \tag{1}$$

where $w_0, w_1, \ldots, w_m$ are the coefficients (weights) of the linear combination, each D_i is a diagonal dilation matrix which scales ψ, each R_i is a rotation matrix, and each $\mathbf{t}_i$ is a translation vector which shifts the input vector $\mathbf{x}$ whose components are $x_1, \ldots, x_n$. The prototype wavelet ψ is obtained by taking the direct product $\psi_1 \psi_2 \cdots \psi_n$ of identical scalar wavelets along each dimension of the input space.

Experiments on target functions of one and two variables support the claim that the wavelet network can achieve an approximation comparable to a wavelet frame using substantially fewer basis function units, and that this approximation is superior to that obtained by a network with the same number of sigmoid units (Zhang & Benveniste, 1992). The curse of dimensionality has not been completely eliminated, however. As the number of input variables increases, the number of combinations of distinctly initialized translations and dilations recommended by Zhang & Benveniste grows combinatorially.

Our present purpose is to investigate the benchmark performance of the comparatively simple architecture obtained if the prototype basis function is a *scalar* wavelet defined in n-space rather than a direct product of n scalar wavelets.

There are two potential advantages in using scalar wavelets rather than multidimensional wavelets for function approximation. The first is that the number of basis functions required does not necessarily grow exponentially with the number of dimensions of the input space, as in the initial basis function placements of Zhang & Benveniste. Intutively, we are attempting to cover an n-dimensional space with semi-local (Hartman & Keeler, 1991) rather than purely local functions.

A second potential advantage of using scalar rather than n-dimensional wavelets is a substantial simplification in Zhang & Benveniste's gradient-descent training algorithm. It is trivial to observe that if the prototype wavelet ψ is a function of one input variable only, say x_1, then only the first row of the matrix product $D_i R_i$ in equation (1) is significant, and hence any possible dilation, rotation, and translation $\psi(D_i R_i(\mathbf{x}-\mathbf{t}_i))$ can be more simply expressed $\psi(v_{i0}+\mathbf{v}_i^T\mathbf{x})$, where $\mathbf{v}_i^T = (v_{i1}, \ldots, v_{in})$ is the first row of the matrix product $D_i R_i$, and $v_{i0} = -\mathbf{v}_i^T\mathbf{t}_i$ serves as an input bias term. Equation (1) can then be more simply expressed in a conventional 2-layers-of-weights form

$$f(\mathbf{x}) \approx g(\mathbf{x}) = w_0 + \sum_{i=1}^{m} w_i \psi(v_{i0} + \mathbf{v}_i^T\mathbf{x}), \tag{2}$$

with ψ now termed an "activation function" or "transfer function." Hence ordinary back propagation, or any of its variants, can be used for training the network of equation (1) when the basis function ψ is a function of only one variable in n-space.

Note that the universal approximation property shown by Zhang & Benveniste (defined in terms of the family of approximations of equation (1) being dense in the set of all square-integrable functions) does not apply to scalar wavelets in n-space, since the composition of a scalar wavelet with a projection from $\Re^n$ to $\Re$ will not in general be a wavelet in $\Re^n$. A similar universal approximation property for scalar wavelets instead follows from the result (Leshno et al., 1993) that for any non-polynomial scalar activation function ψ, the family of approximations of equation (2) is dense in the set of all continuous functions over $\Re^n$.

The potential advantage of scalar wavelets in comparison with conventional sigmoid basis functions is that scalar wavelets are more localized. This could potentially lead to faster training with only a portion of the basis function units participating significantly in the approximation of each training example. To see whether this potential advantage can in fact be realized, networks of scalar wavelet and sigmoid basis functions were trained on the same benchmark problem using the same algorithm with the same parameter settings. The basis functions were then compared both in terms of training speed and in terms of generalization performance after a fixed amount of training.

3. PROCEDURE

Prediction of the Mackey-Glass chaotic time series (Mackey & Glass, 1977) is a widely-recognized benchmark for comparing neural network architectures. The specific time series employed in this study, and by many previous investigators, (Farmer & Sidorowich, 1987; Moody & Darken, 1989; Tenorio, 1990; Hartman & Keeler, 1991; Martinetz et al., 1993) is the time series generated by the differential equation

$$\frac{dx(t)}{dt} = -bx(t) + a \cdot \frac{x(t-T)}{1 + [x(t-T)]^{10}} \tag{3}$$

for $a = 0.2$, $b = 0.1$, and $T = 17$. The prediction of time series similar to this arises in physiological phenomena such as clinically-observed breathing irregularities and oscillations in white blood cell counts (Mackey & Glass, 1977). The benchmark task typically associated with this time series is to predict the value of the time series at time $t + 85$ from the four values at times t, $t - 6$, $t - 12$, and $t - 18$. Following previous work, the training set consisted of 500 randomly selected samples from the segment of the time series between $t = 500$ and $t = 4000$, and the test set consisted of 500 consecutive samples beginning at $t = 4000$. All data were scaled based on the mean and standard deviation of the training data.

The quickprop algorithm (Fahlman, 1988), a heuristic second-order gradient descent method, was employed because of its speed and robustness. The same algorithm and parameter settings were used for training comparable $1\frac{1}{2}$ - layer sigmoid networks and $1\frac{1}{2}$ - layer scalar wavelet networks, the only difference between the two being the computation of the activation function (transfer function) and its derivative. In the sigmoid network, each node in the first layer computes a symmetric version of the conventional sigmoidal transfer function , $\psi(x) = 1/(1 + e^{-x}) - \frac{1}{2}$. The scalar basis function network employs instead the "Gaussian derivative" wavelet $\psi(x) = -xe^{-x^2/2}$ as its transfer function. In both cases, the second "half-layer" of the $1\frac{1}{2}$ - layer network consists of a single output node with the identity transfer function.

The two basis functions were compared in networks of 16, 32, 64, and 128 basis function units. For uniform comparisons over this range of network sizes, the learning rate ϵ of the quickprop algorithm was always set at 0.09. This may be considered a conservative approach to demonstrating the performance of the scalar wavelet network, since in many of the random initializations studied, learning rates higher than this caused the sigmoid network to diverge, while the scalar wavelet network still converged, at a more rapid rate.

Each training run consisted of 1 million passes over the training set, using the conventional mean-squared error function, and using the default values in the public-domain implementation of the quickprop algorithm for all other parameters (namely, a maximum growth factor μ of 1.75, an initial random weight range of -0.7 to +0.7, a derivative-offset of 0.1, and a weight decay of 0.0001). For each size and type of network, the result reported below is the best obtained by restarting each training run 5 times with the weights reinitialized using a different random seed (for a total of 5 million passes over the set of training data).

4. RESULTS AND DISCUSSION

Figure 1 compares the training progress of sigmoid and scalar wavelet networks for each network size. The normalized error over the training data is plotted as a function of the training passes over this data set. (Normalized error is defined in the usual way as the rms prediction error divided by the standard deviation of the set of correct predictions.) For each network size, throughout the training process, the quickprop algorithm achieved a lower error for scalar wavelet units than for sigmoid units.

The two solid lines of Figure 2 compare the generalization performance of the resulting trained sigmoid and scalar wavelet networks. Generalization performance is measured as the normalized error in the predictions made by each trained network for the test data set described in the preceding section. For all network sizes studied, scalar wavelet networks appear to achieve better generalization performance than comparable networks of sigmoidal units.

For comparison, Figure 2 also shows results for this time series prediction benchmark using other methods unrelated to the present study. The dashed line of Figure 2 shows results from our implementation of the K-means clustering algorithm (Moody

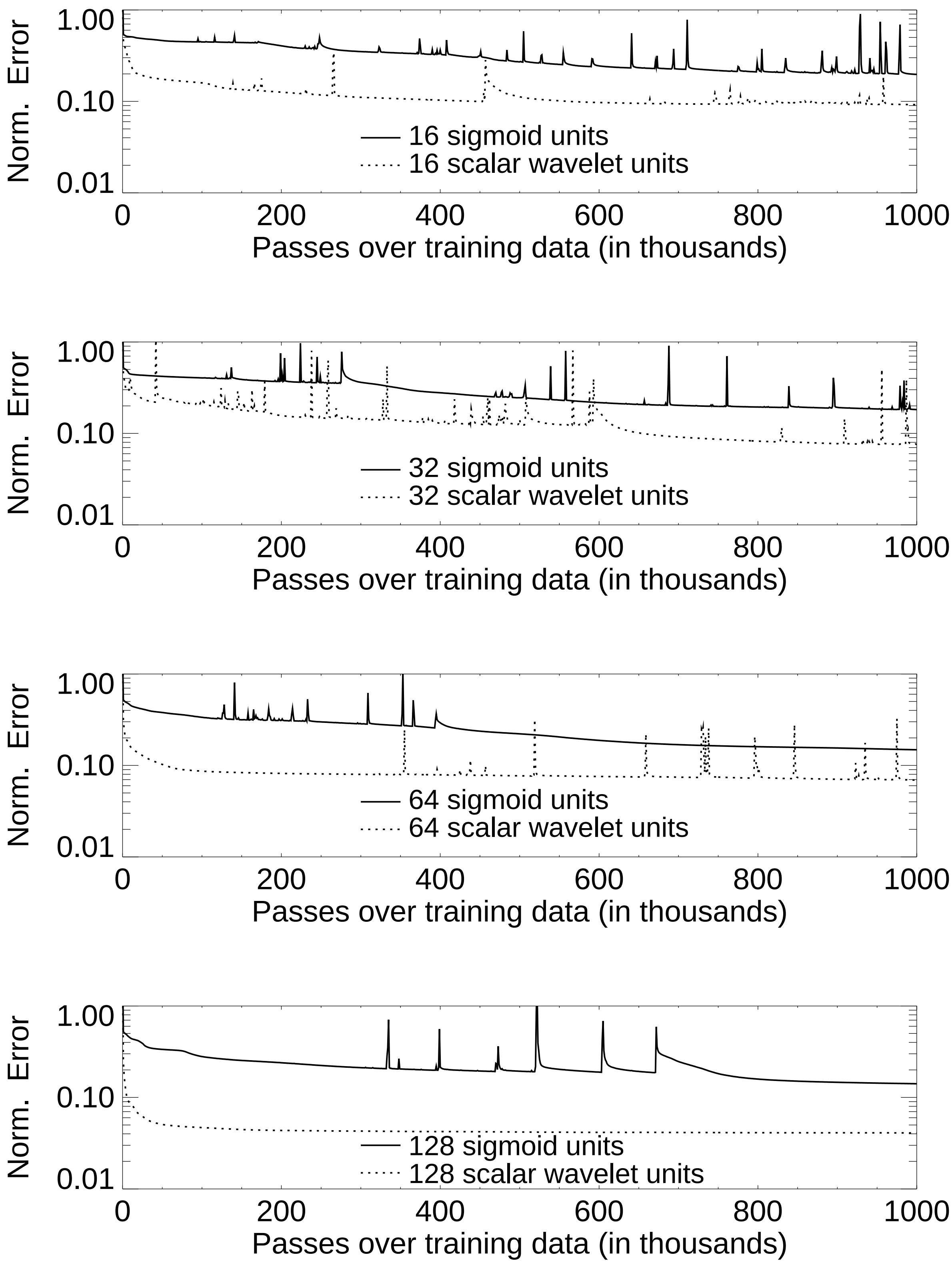

Fig. 1: Error reduction during training of $1\frac{1}{2}$ - layer networks employing scalar wavelet (dotted line) versus sigmoid (solid line) basis function units. Both cases use the quickprop gradient descent algorithm with the learning rate ϵ set to 0.09.

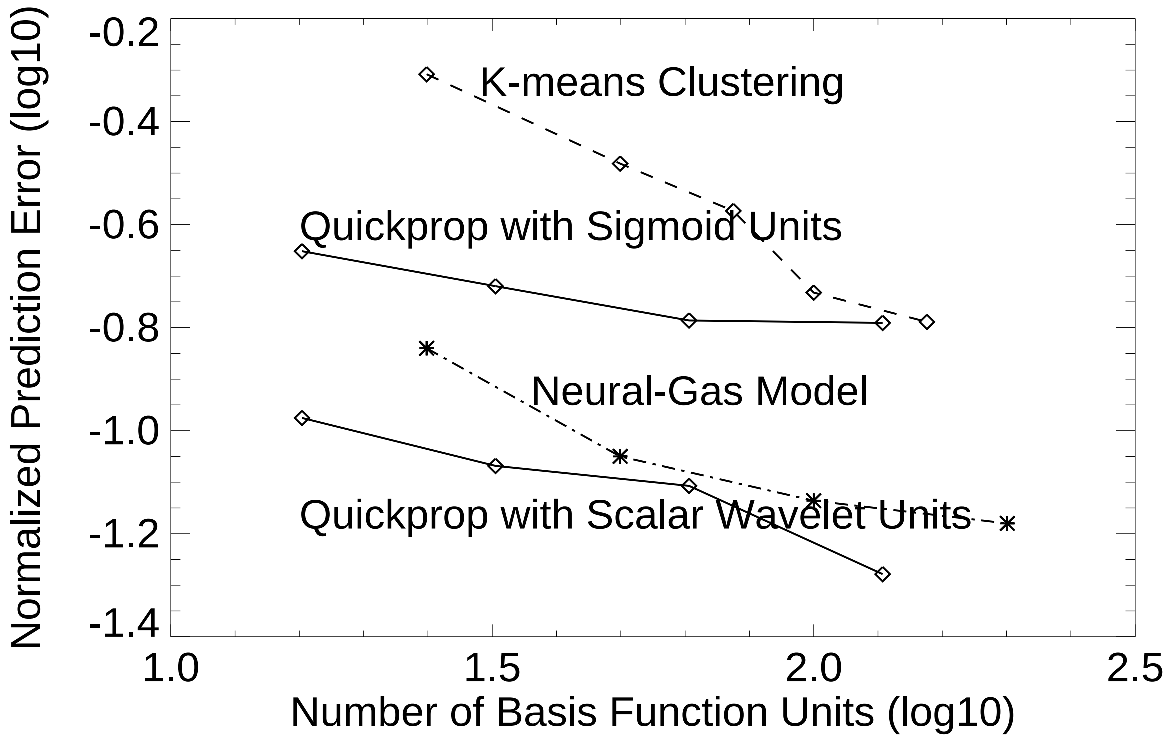

Fig. 2: The solid lines show the normalized prediction errors for the trained networks of sigmoid units and scalar wavelet units resulting from Figure 1. In both cases, these prediction errors were measured on the test data set after 5 million passes over the training data set. For reference, results for this same benchmark from the K-means clustering model (Moody & Darken, 1989) and the neural-gas model (Martinetz et al., 1993) are also shown. (The data points plotted with asterisks are taken from Figure 5 of Martinetz et al.)

& Darken, 1989). These results are consistent with the linear (in log-log coordinates) relationship observed for larger numbers of units by Moody & Darken. The dash-dotted line shows results presented for the neural-gas model (Martinetz et al., 1993) under comparable conditions.

Based on these initial comparisons on a standard benchmark, the scalar wavelet network appears promising as a candidate for further evaluation in the large number of practical applications which currently employ back propagation methods based on sigmoidal units.

ACKNOWLEDGMENT

The implementation of the quickprop algorithm was based on Terry Regier's C-language translation of Scott Fahlman's quickprop program, with modifications by N. Karunanithi. I thank these people for making their source code freely available.

REFERENCES

1. Bakshi, B. R. & Stephanopoulos, G. (1992). Wavelets as basis functions for localized learning in a multi-resolution hierarchy. In *International Joint Conference on Neural Networks*, pages II–140–145. IEEE.

2. Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005.

3. Fahlman, S. E. (1988). Faster-learning variations on back-propagation: An empirical study. In *Proceedings of the 1988 Connectionist Models Summer School*. Morgan Kaufmann.

4. Farmer, J. D. & Sidorowich, J. J. (1987). Predicting chaotic time series. *Physical Review Letters*, 59(8):845–848.

5. Hartman, E. & Keeler, J. D. (1991). Predicting the future: Advantages of semilocal units. *Neural Computation*, 3:566–578.

6. Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynominal activation function can approximate any function. *Neural Networks*, 6:861–867.

7. Mackey, M. C. & Glass, L. (1977). Oscillation and chaos in physiological control systems. *Science*, 197:287–289.

8. Martinetz, T. M., Berkovich, S. G., & Schulten, K. J. (1993). Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569.

9. Meyer, Y. (1992). *Wavelets and Applications*. Springer-Verlag, New York.

10. Moody, J. & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294.

11. Pati, Y. C. & Krishnaprasad, P. S. (1993). Analysis and synthesis of feedforward neural networks using discrete affine wavelet transformations. *IEEE Transactions on Neural Networks*, 4(1):73–85.

12. Tenorio, M. F. (1990). Self-organizing network for optimum supervised learning. *IEEE Transactions on Neural Networks*, 1(1):100–110.

13. Zhang, Q. & Benveniste, A. (1992). Wavelet networks. *IEEE Transactions on Neural Networks*, 3(6):889–898.