# COMPARISON OF TWO LEARNING NETWORKS FOR TIME SERIES PREDICTION

**Daniel Nikovski**
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213, USA
Email: Daniel.Nikovski@cs.cmu.edu

**Mehdi Zargham**
Department of Computer Science
Southern Illinois University at Carbondale
Carbondale, Illinois 62901, USA
Email: mehdi@cs.siu.edu

## ABSTRACT

Hierarchical mixtures of experts (HME) [JJ94] and radial basis function (RBF) networks [PG89] are two architectures that learn much faster than multilayer perceptrons. Their faster learning is due not to higher-order search mechanisms, but to restricting the hypothesis space of the learner by constraining some of the layers of the network to use linear processing units. It can be conjectured that since their hypothesis space is restricted in the same manner, the approximation abilities of the two networks should be similar, even though their computational mechanisms are different. An empirical verification of this conjecture is presented, based on the task of predicting a nonlinear chaotic time series generated by an infrared laser.

## INTRODUCTION

The problem of predicting the continuation of a given time series is fundamentally a machine learning problem, since such a continuation requires the extraction by the prediction method of the rules that govern the dynamics of the time series. This extraction has to be done from examples of past values of the time series. The use of past values of the time series for reconstructing the dynamics of the underlying dynamic system is justified by Takens' theorem [WG94].

If the dynamic system that generates the time series is linear, two suitable forms for representing the inferred knowledge are regression coefficients and transfer functions (ARIMA models) [LS83]. However, when the underlying dynamic system is nonlinear, other representations are necessary. One possibility is to use locally linear models that store all past values of the time series and then produce a prediction based on locally linear interpolation [GW94]. While giving good predictions, such models are obviously useless when the time series is of infinite duration.

The other possibility is to use learning methods such as multilayer perceptrons (MLP) or other learning networks that concentrate the extracted knowledge in a small number of parameters, which are adjusted by on-line learning rules [GW94]. The requirements to such learning networks are that they should be universal function approximators, have good generalization power, be able to learn quickly from examples, and be implementable in hardware. The most popular learning networks, MLP, possess all of the above features except fast learning; convergence of MLP for nontrivial problems is in the order of hundreds of thousands of iterations and can be impractical for most real-time engineering applications even if it is accelerated by hardware.

The slow learning in MLP has motivated the development of alternative learning architectures that can learn faster. Two such learning models are radial basis functions [PG89] and hierarchical mixtures of experts [JJ94]. The two models are much faster than MLP because a significant part of their processing is linear, though in a different way. Both methods have

been used for time series prediction — RBF by Casdagli [Cas89] and HME by Waterhouse and Robinson [WR95].

The approximation power and convergence time of each of the two methods have been compared with MLP so far [Cas89, JJ94], but on different test problems. It is thus difficult to see which of the two methods is better for time series prediction or any other nonlinear identification task, and it is of practical interest to compare them on a single test problem. Section 2 describes the chosen problem and the testing procedure. The architecture and performance on the test problem of RBF and HME are presented in sections 3 and 4 respectively. Section 5 provides a statistical comparison of the two methods, and section 6 discusses the results and concludes.

## DESCRIPTION OF THE PROBLEM

Benchmarking of nonlinear time series prediction methods requires test problems that are representative of the problems commonly encountered in practice. Such nonlinear time series have been provided by the organizers of the Santa Fe Institute competition in time series prediction [GW94]. We used as test problem one of these time series, consisting of output of a $CH_3$ far-infrared laser. This nonlinear chaotic time series has been predicted successfully with a locally linear model by Sauer [Sau94] and with a time-delay neural network by Wan [Wan94]. Wan used an embedding dimension of 8; this value for the dimensionality of the embedding space was adopted in our experiments too, which means that our networks had 8 inputs each. The original data were measured as integers in the range from 0 to 255, which introduces quantization error of about 0.2%.

Based on the provided time series data, one training and nine testing data sets were prepared. The training set had 992 examples, and the testing sets had 1000 examples each. RBF and HME used the training set to estimate their parameters. The first of the testing sets was used for cross-validation - i.e., the estimated model that produced best out-of-sample error on the first testing set was assumed to be the optimal one for the respective architecture. The out-of-sample error of this model was tested on the remaining eight testing sets in order to compare the two learning methods. All errors reported in the paper are normalized root mean squared errors (NRMSE). For the purpose of comparison with linear prediction methods, a recursive least squares (RLS) autoregressive model was tested too [LS83].

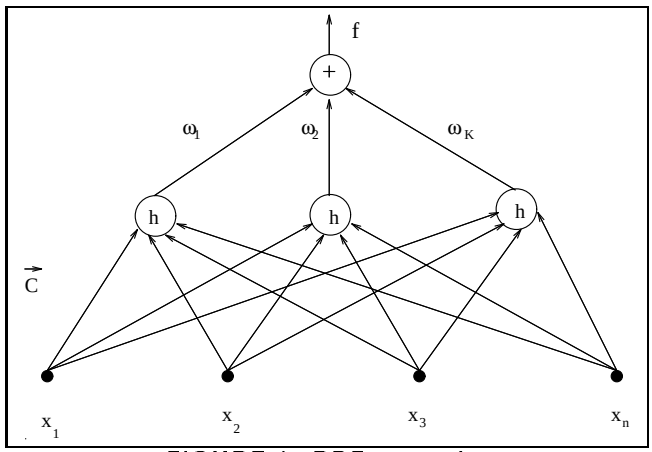


**FIGURE 1. RBF network**

## PREDICTION WITH RADIAL BASIS FUNCTIONS

Radial Basis Functions (RBF) have been known in approximation theory as a powerful and computationally efficient method for interpolation and approximation of functions [PG89]. With the development of the connectionist approach to computation, it was noticed that the computation performed by these functions can be carried out by parallel and distributed processing elements similar to multilayer perceptrons [PG89].

An RBF network consists of a number of units whose output is a nonlinear function $h$ of the unit's input (Fig. 1).

The output of the network $y$ is given by

$$y = f(\mathbf{x}) = \sum_{\alpha=1}^{K} \omega_\alpha h(\| \mathbf{x} - C_\alpha \|)$$

where $\mathbf{x}$ is the vector input to the network, $\omega_\alpha$ are the weights that should be estimated, $C_\alpha$ are vectors called knots, and $\| \cdot \|$ is the $L^2$ norm on the space of input vectors. In our experiments the knots $C_\alpha$ were chosen to be a subset of the training examples; other possible approaches include various unsupervised clustering schemes or supervised adjustment of the knots. The function $h$ belongs to the class of radial basis functions [PG89]. For our experiments, the linear function $h(r) = r$ was used.

The estimation of the weights $\omega_\alpha$ is a generalized linear squares problem and is solvable in time $O(K^3)$ in the number of knots $K$. A plot of the in- and out-of-sample NRMSE versus the number of knots is given in Fig. 2 for the problem of laser output prediction.

The in-sample error decreases gradually to zero, as expected. The out-of-sample error follows closely, but

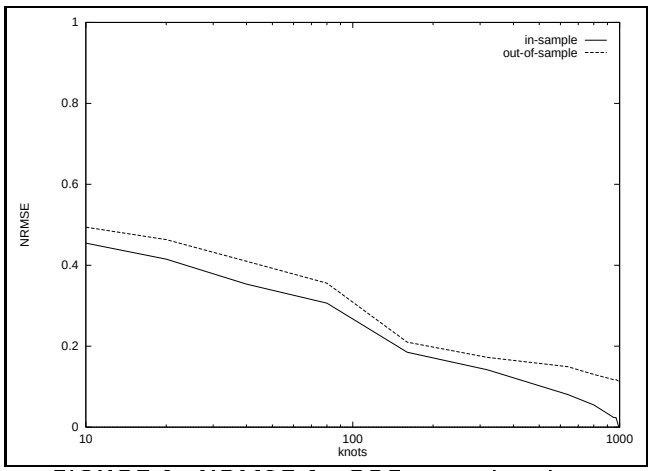


**FIGURE 2. NRMSE for RBF approximation**

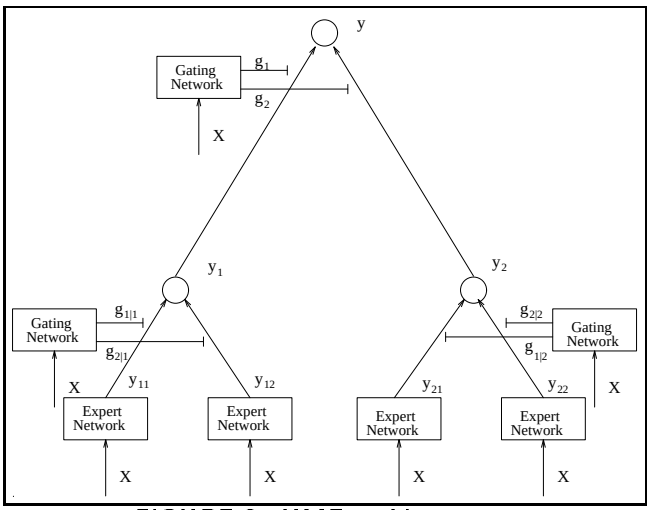


**FIGURE 3. HME architecture**

flattens to 0.10 when the number of knots approaches the number of training examples. Some overfitting occurs; the lowest out-of-sample error is attained for 990 knots, 2 less the number of training examples. Consequently, the optimal model from cross-validation with the first testing set is the model with 990 knots.

This model was tested on the remaining 8 testing sets, resulting in average out-of-sample error $\bar{E}_{RBF} = 0.10212$.

## PREDICTION WITH HIERARCHICAL MIXTURES OF EXPERTS

The HME architecture is a general nonlinear function approximation model introduced by Jordan and Jacobs [JJ94]. In particular, it can be used for nonlinear regression in time series prediction. The architecture consists of a tree-like hierarchy of gating nodes and a set of experts at the leaves of the tree (Fig. 3).

If HME is to be used for nonlinear regression, each of the experts forms a linear prediction $\hat{y}_{ij}$ of the true output of the network $y$ for a given input vector $\mathbf{x}$:

$$\hat{y}_{ij} = \mathbf{u}_{ij}^T \mathbf{x}$$

The purpose of the gating nodes is to assign weights to the predictions $\hat{y}_{ij}$ of the individual experts - these weights should be proportional to the precision with which an expert approximates the function at a particular location. For this purpose, the weights depend on the input $\mathbf{x}$ too. For the case of binary trees, shown in Fig. 3, two values $g_1$ and $g_2$ are output by the gate at the root of the tree, with their sum equal to 1:

$$g_i = \frac{e^{\xi_i}}{\sum_k e^{\xi_k}},\ i = 1, 2,\ k = 1, 2$$

$$\xi_i = \mathbf{v}_i^T \mathbf{x}$$

Here the vector $\mathbf{v}_i$ parametrizes the gating expert. The predicted output $\hat{y}$ is then a weighted average of the predictions at the lower level $\hat{y}_i$:

$$\hat{y} = \sum_{i=1}^{2} g_i \hat{y}_i$$

The weights of the gating node at the root of the tree perform a partitioning of the input space into two parts, in each of which one of the weights is close to 1, while the other weight is close to 0. In this way a "soft split" is formed, whose steepness and orientation are determined by the vector $\mathbf{v}_1 - \mathbf{v}_2$. The greater the length of this vector, the sharper the split is. The splitting continues recursively on the lower levels of the tree.

The predictions at the lower level are formed in a similar way, this time using directly the predictions of the experts:

$$\hat{y}_i = \sum_{j=1}^{2} g_{j|i} \hat{y}_{ij}$$

Jordan and Jacobs [JJ94] derived learning rules for the experts and the gates on the basis of a probability model and a maximum-likelihood approach. The probability distribution of the output of a particular linear expert is given by

$$P(y|\mathbf{x}, \mathbf{u}_{ij}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\hat{y}_{ij})^2}{2}}$$

By differentiating the likelihood that the observed output $y$ is generated by the HME model with a partic-

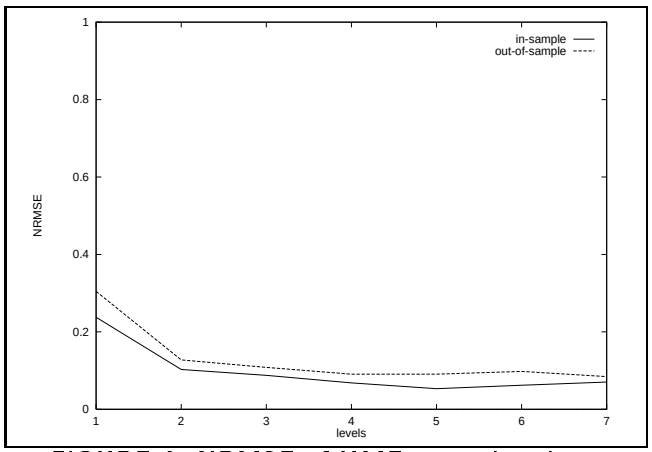


**FIGURE 4. NRMSE of HME approximation**

ular set of parameters $\theta$, the following on-line gradient ascent learning rules result:

$$\Delta \mathbf{u}_{ij} = \rho h_i h_{j|i} (y - \hat{y}_{ij}) \mathbf{x}$$

$$\Delta \mathbf{v}_i = \rho (h_i - g_i) \mathbf{x}$$

$$\Delta \mathbf{v}_{ij} = \rho h_i (h_{j|i} - g_{j|i}) \mathbf{x}$$

where $\rho$ is a learning rate and the quantities $h_i$, $h_{j|i}$ and $h_{ij}$ are posterior probabilities defined as follows:

$$h_i = \frac{g_i \sum_j g_{j|i} P_{ij}(y|\mathbf{x}, \theta)}{\sum_i g_i \sum_j g_{j|i} P_{ij}(y|\mathbf{x}, \theta)}$$

$$h_{j|i} = \frac{g_{j|i} P_{ij}(y|\mathbf{x}, \theta)}{\sum_j g_{j|i} P_{ij}(y|\mathbf{x}, \theta)}$$

$$h_{ij} = \frac{g_i g_{j|i} P_{ij}(y|\mathbf{x}, \theta)}{\sum_i g_i \sum_j g_{j|i} P_{ij}(y|\mathbf{x}, \theta)}$$

Jordan and Jacobs have also derived another set of learning rules, based on the Expectation-Maximization (EM) developed in statistics [JJ94]. The EM learning rules learn faster than the ones based on the maximum-likelihood approach.

In the HME architecture, the number of estimated parameters and hence the achieved generalization depend on the number of levels in the hierarchy. For the laser test problem, the number of levels was varied and the in-sample and out-of-sample errors were plotted, similarly to the testing with RBF networks. The results are shown in Fig. 4. The tree was binary and the learning rate $\rho$ was set to 0.01. For each model, 10000 iterations were performed.

The least out-of-sample error was attained for a seven-level network (128 experts). This network was used for comparison on the remaining 8 testing sets. Notice that this was not the network with smallest in-sample error.

## COMPARISON OF THE TWO LEARNING MODELS

The relative performance of RLS and the best RBF and HME models is shown in Table 1.

While the superiority of HME and RBF over RLS is obvious, it is not clear how the RBF and HME networks compare with each other. To this end a paired two-tailed t-test was performed on the 8 testing sets to verify the hypothesis that one of the two networks is superior. The results are shown in Table 2.

Table 2 shows that there is no statistical significance to the hypothesis that RBF and HME have different performance on the learning task at hand.

## CONCLUSIONS

The experiments presented in the previous sections demonstrate that the HME and RBF architectures possess comparable approximation abilities on a non-linear identification task. While it might be difficult to show analytically such similarity in the general case, the presented empirical results suggest that the two learning networks consider similar hypothesis spaces (or, in other words, impose similar smoothness constraints on their approximations.) This observation can be used to infer expectations about the performance on a given task of one of the networks, knowing the approximation error of the other. One application might be to train fast an RBF network and if

**Table 1. NRMSE of RLS, RBF, and HME on the laser output time series.**

| | RLS | RBF | HME |
|---|---|---|---|
| In-sample | 0.43309 | 0.00247 | 0.07028 |
| Out-of-sample ($\bar{E}$) | 0.44586 | 0.10212 | 0.10414 |

**Table 2. Comparison of RBF, HME, and RLS: t-statistic and attained significance level $p$.**

| | $t$ | $p$ |
|---|---|---|
| RBF vs. HME | 0.355286 | 0.440666 |
| RBF vs. RLS | 42.07407 | $< 10^{-11}$ |
| HME vs. RLS | 42.73772 | $< 10^{-11}$ |

the performance is satisfactory, proceed with the more time-consuming training of a HME network, whose approximation is piecewise linear and because of that can be more readily used for prediction and control applications.

## REFERENCES


[Cas89] Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D*, vol. 35, 335-356.

[GW94] Gershenfeld, N.A., and A.S. Weigend (1994). The future of time series: learning and understanding. In Weigend, A.S, and N.A. Gershenfeld (Eds.) (1994) *Time Series Prediction.* Reading, MA: Addison-Wesley, 1-70.

[JJ94] Jordan, M.I., and Jacobs, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, vol. 6, 181-214.

[LS83] Ljung, L., and T. Söderström (1983). *Theory and Practice of Recursive Identification.* Cambridge, MA: MIT Press.

[PG89] Poggio, T., and F. Girosi (1989). *A Theory of Networks for Approximation and Learning.* AI Memo 1140, MIT AI Lab.

[Sau94] Sauer, T. (1994). Time series prediction by using delay coordinate embedding. In Weigend, A.S, and N.A. Gershenfeld (Eds.) (1994) *Time Series Prediction.* Reading, MA: Addison-Wesley, 175-193.

[Wan94] Wan, E.A.(1994). Time series prediction by using a connectionist network with internal delay lines. In Weigend, A.S, and N.A. Gershenfeld (Eds.) (1994) *Time Series Prediction.* Reading, MA: Addison-Wesley, 195-217.

[WR95] Waterhouse, S.R., and A.J. Robinson (1995). Nonlinear prediction of acoustic vectors using hierarchical mixtures of experts, in G. Tesauro, D.S. Touretzky, and T.K. Leen, (Eds.), *Neural Information Processing Systems 7*, Cambridge, MA: MIT Press.

[WG94] Weigend, A.S, and N.A. Gershenfeld (Eds.) (1994) *Time Series Prediction.* Reading, MA: Addison-Wesley.