RATES OF CONVERGENCE OF THE RECURSIVE RADIAL BASIS FUNCTION NETWORKS

J. Mazurek[1] *A. Krzyżak*[2] * *A. Cichocki*[3]

[1]Neurolab GmbH, Germany, email jama@nws.e-technik.tu-muenchen.de
[2]Dept. of Computer Science, Concordia University, Canada, email krzyżak@cs.condordia.ca
[3]FRP Riken, Lab. for Artificial Brain Systems, Wako-city, Japan, email cia@zoo.riken.go.jp

ABSTRACT

Recursive radial basis function (RRBF) neural networks are introduced and discussed. We study in detail the nets with diagonal receptive field matrices. Parameters of the networks are learned by a simple procedure. Convergence and the rates of convergence of RRBF nets in the mean integrated absolute error (MIAE) sense are studied under mild conditions imposed on some of the network parameters. Obtained results give also upper bounds on the performance of RRBF nets learned by minimizing empirical L_1 error.

1. INTRODUCTION

A large number of the multilayer feedforward networks described in the literature consist of units that compute an inner product of a weight vector and input vector followed by a nonlinear activation function (e.g. sigmoidal function), see e.g. Cichocki and Unbehauen [3]. However recently a number of authors have discussed the use of processing units that compute a distance measure between an input vector and a weight vector, usually followed by a Gaussian shaped function. Radial basis function (RBF) nets are examples of such networks. RBF net contains only one hidden layer with processing nodes which realize the radial basis function. Furthermore, the activation functions are usually nonmonotonic and local. The output units perform simple linearly weighted summation of its inputs. A number of theoretical results on Radial Basis Function (RBF) networks have been obtained, see Xu, Krzyżak and Yuille [17] for a long list of references. It has been shown that RBF nets can be naturally derived from the *regularization theory* (Poggio and Girosi [13], and that RBF nets have the universal approximation ability (Hartman, Keeler and Kowalski [7], Park and Sandberg [12]) as well as the so-called best approximation ability (Girosi et al [5]). Specht [15] introduced *probabilistic neural networks* and pointed out the connection between RBF nets and *Parzen window* estimators of probability density [14]. Xu et al [17] found the connection between RBF nets and *kernel regression estimate* [6] and studied universal convergence and upper bounds on the rates of convergence of RBF nets. Rates of convergence of RBF nets approximation error were studied by Girosi and Anzellotti [4] and the rates of estimation error are given by Niogy and Girosi [11].

Most theoretical studies on RBF nets were limited to non-recursive versions in which radial functions were identical at each hidden node. Recently Krzyżak and Linder [8] considered general recursive RBF nets (with general receptive field matrix). Niogy and Girosi [11] considered the recursive RBF nets with receptive field matrix being the identity matrix. In both papers the learning process was carried out by computationally intensive minimization of the empirical L_2 error. In the present paper we consider recursive RBF nets (RRBF nets) with diagonal receptive field matrices. These nets are fairly simple but also very flexible and sufficient for most applications. The nets are trained by a simple procedure randomly selecting centers and output weights from the training sequence. The performance of the nets is measured by the mean integrated absolute error (MIAE) which is important measure in robust estimation. We study generalization ability of RRBF nets together with convergence and the rates of convergence. Our results provide also upper bound on the performance of general RRBF with positively defined receptive field matrices and with parameters learned by minimization of the empirical L_1 error.

2. RBF AND RRBF NETS

Let (X, Y) be a pair of random vectors in $R^d \times R^m$ and $R(x) = E\{Y|X = x\} = [r^{(1)}(x), \cdots, r^{(m)}(x)]^T$ be the corresponding regression function. Let μ denote the probability measure of X. Consider a network $f_{n,N}(x)$ learned by a training set $D_N = \{X_i, Y_i\}_1^N$, where N is the number of training samples and n is the size of the network, e.g. the number of hidden neurons in the network. Two types of RBF nets are prevalent in the literature:

- standard nets [5, 8, 12]

$$g_n(x) = \sum_{i=1}^{n} w_i K([x - c_i]^t \Sigma^{-1} [x - c_i]) \qquad (1)$$

- normalized nets [10, 17]

$$g_n(x) = \frac{\sum_{i=1}^{n} w_i K([x - c_i]^t \Sigma^{-1} [x - c_i])}{\sum_{i=1}^{n} K([x - c_i]^t \Sigma^{-1} [x - c_i])} \qquad (2)$$

where $K(r^2)$ is a radial basis function, $c_i, i = 1, \ldots, n$ are the center vectors, $w_i, i = 1, \ldots, n$ are the weight vectors and Σ is arbitrary $d \times d$ positive definite matrix which

*This work is supported by NSERC grant A0270, FCAR grant EQ-2904 and by the Alexander von Humboldt Foundation of Germany.

controls the receptive field of the basis functions. The most common choice for $K(r^2)$ is the Gaussian function, $K(r^2) = e^{-r^2}$ with $\Sigma = \mathrm{diag}(\sigma_1(n)^2, \ldots, \sigma_d(n)^2)$, but other functions have also been used (see [13] for other choices). Networks (1) are related to Parzen density estimate

$$p_n(x) = p_n(x, \mathcal{D}_n^g) = \frac{1}{nh_n^d} \sum_{i=1}^{n} \phi(\frac{x - X_i}{h_n})$$

where ϕ is the normalized kernel and h_n is a bandwidth (see Scott [14] and references therein), and to so called *probabilistic neural network* proposed by Specht [15]. Networks (2) are related to the kernel regression estimate [6, 17]

$$R_n(x) = \frac{\sum_{i=1}^{n} Y_i \phi(\frac{x - X_i}{h_n})}{\sum_{i=1}^{n} \phi(\frac{x - X_i}{h_n})}$$

which is the weighted average of Y_i which approximates conditional mean of the output given input $E(Y|X = x)$ with adjustable weights nonlinearly depending on the input observations and x.

In the present paper we consider the recursive version of (2)

$$f_n(x) = \frac{\sum_{i=1}^{n} w_i K([x - c_i]^t \Sigma_i^{-1} [x - c_i])}{\sum_{i=1}^{n} K([x - c_i]^t \Sigma_i^{-1} [x - c_i])} \tag{3}$$

in which all the parameters besides Σ_i are defined as in (2) and the receptive field is a diagonal matrix $\Sigma_i = \mathrm{diag}(\sigma_{i1}^2, \ldots, \sigma_{id}^2)$. To simplify the notation define $K(r) = \phi(r^2)$, $||x_i|| = [\sum_{k=1}^{d} x_{ik}^2]^{1/2}, ||x_i||_{\Sigma^{-1}} = [x_i^t \Sigma^{-1} x_i]^{1/2} = [\sum_{k=1}^{d} (x_{ik}/\sigma_{ik})^2]^{1/2}, x_i = (x_{i1}, \ldots, x_{id})^t$. All the parameters to be learned may be gathered into vector $\theta = (w_1, \ldots, w_n, c_1, \ldots, c_n, \Sigma_1, \ldots, \Sigma_n)$. The following are the possible learning strategies

1. minimize the empirical error with respect to θ (see e.g. [1]), i.e.

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} |Y_i - f_n(X_i)| \to \theta^*. \tag{4}$$

   Denote the resulting optimal net by $f_{n,N}^*$.

2. cluster X_i in $\mathcal{D}_N$ and assign c_i to cluster centers. Remaining parameters are obtained by minimization process in (4)

3. assuming that the size of the learning sequence is larger than the number of nodes in the hidden layer ($N > n$) draw a subset $\mathcal{D}_n = \{X_i, Y_i\}_1^n$ from $\mathcal{D}_N$ and assign $X_i \to c_i, Y_i \to w_i, i = 1, \ldots, n$ and choose Σ_i according to the rules given in the next section.

Of the three strategies described above we choose strategy (3) as the simplest but still yielding convergent RRBF nets (see section 3). Thus network (3) has been reduced to RRBF net

$$f_n(x) = \frac{\sum_{i=1}^{n} Y_i K([x - X_i]^t \Sigma_i^{-1} [x - X_i])}{\sum_{i=1}^{n} K([x - X_i]^t \Sigma_i^{-1} [x - X_i])}. \tag{5}$$

It is clear that $K(||x||_{\Sigma^{-1}})$ is no longer radially symmetric function of x even when $K(x)$ is, since $||x||_{\Sigma^{-1}} = \mathrm{const}$ is an ellipsoid with axes parallel to coordinate axes. Most of the results in the literature were obtained for radially symmetric receptive fields [2, 10, 13], but our convergence results in the next section do not require radially symmetric basis functions.

The performance of network (5) can be measured by either

$$E|R(X) - f_n(X)| \tag{6}$$

or

$$E|R(X_1) - f_n(X_1)| \tag{7}$$

where X in (6) is independent of $\mathcal{D}_N$ (generalization) and X_1 in (7) is the first measurement in $\mathcal{D}_N$ (no generalization). We consider index (7) since we can bound the performance of $f_{n,N}^*$ by (7)

$$E|Y_1 - g_{n,N}^*(X_1)| \le E|Y_1 - f_n(X_1)|$$

when learning strategy 1. is used (this is MIAE analog of Lemma 1 in [17]). Since convergence analysis of index (6) easily follows from analysis of (7) the convergence analysis in the next section is confined to (7).

3. CONVERGENCE AND RATES OF RRBF NETS

In this section we study asymptotic behavior of RRBF nets. The next theorem gives sufficient conditions for convergence of net (5) when the size of the learning sequence increases without restrictions.

Theorem 1 (RRBF convergence) *Let* $E|Y| < \infty$,

$$c_1 I_{S_{0,r}} \le K(x) \le c_2 I_{S_{0,R}}, \quad 0 < r < R < \infty, c_1, c_2 > 0 \tag{8}$$

and assume

$$n \prod_{i=1}^{d} \sigma_i \to \infty$$
$$\limsup_n \frac{\sum_{i=1}^{n} \prod_{k=1}^{d} \sigma_{ik}}{n \prod_{k=1}^{d} \sigma_k} = \gamma < \infty$$
$$\frac{\sum_{i=1}^{n} \prod_{k=1}^{d} \sigma_{ik} I_{\{||\Sigma_i|| \ge \epsilon\}}}{n \prod_{k=1}^{d} \sigma_k} \to 0$$

as $n \to \infty$,
where I_A *denotes indicator of set* A, $S_{x,r} = \{y : ||y - x|| \le r\}$, $\sigma_k = \min_{1 \le i \le n} \sigma_{ik}, k = 1, \cdots, d$ *and* $||\cdot||$ *is an Euclidean matrix or vector norm. Then*

$$E|R(X_1) - f_n(X_1)| \to 0$$

as $n \to \infty$.

In Theorem 1 a natural condition $E|Y| < \infty$ is imposed on the output. Assumption (8) is satisfied for arbitrary finite kernels compactly supported and bounded away from zero at the origin.

Theorem 2 (RRBF convergence rate) *Let μ denote the probability measure of X with a compact support, $E|Y|^{1+s} < \infty \quad s > 0$ and*

$$c_1 I_{S_{0,r}} \leq K(x) \leq c_2 I_{S_{0,R}}, \quad 0 < r < R < \infty, \quad c_1, c_2 > 0$$
$$\overline{\sigma} \to 0, \quad n^{s/(s+2)} \prod_{i=1}^{d} \sigma_i \to \infty$$
$$as \quad n \to \infty.$$

Also let R satisfy Lipschitz condition

$$|R(x) - R(y)| \leq \beta ||x - y||^{\alpha}, \quad 0 < \alpha \leq 1, \quad \beta > 0.$$

Then

$$E|R(X_1) - f_n(X_1)|$$
$$= O\left(\max\left(\frac{1}{\sqrt{n^{s/(2+s)}\overline{\sigma}}}, \frac{\sum_{i=1}^{n} ||\Sigma_i||^{\alpha} \prod_{k=1}^{d} \sigma_{ik}}{n\overline{\sigma}}\right)\right)$$

where $\overline{\sigma} = \prod_{k=1}^{d} \sigma_k$.

When Σ_i have all diagonal elements identical then the MIAE convergence rate above becomes $O(n^{-\frac{\alpha s}{(2+s)(2\alpha+d)}})$.

4. SIMULATION RESULTS

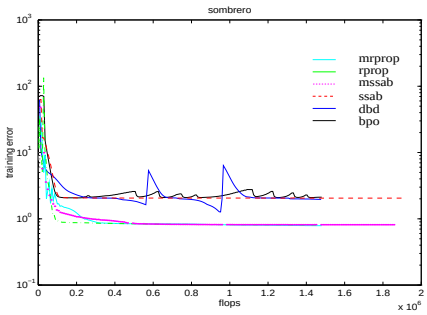
In Figures 1-3 we show the exemplary simulations results in application to the function approximation problem by the standard and generalized RBF networks. We use several adaptive learning algorithms and several radial functions, and we learn output weights, centers and covariance matrices by minimizing empirical L^1 and L^2 errors using stochastic gradient descent approach [3]. We tested the algorithms, e.g. on the folowing 2D functions: wave $f(x_1, x_2) = x_1 \exp[-(x_1^2 + (x_2/0.75)^2)]$, sombrero $f(x_1, x_2) = \sin(\sqrt{x_1^2 + x_2^2})/\sqrt{x_1^2 + x_2^2}$ and other well-known data-set benchmarks. We have extensively investigated and compared various network architectures: SRBF (Standard RBF), ERBF (Eliptic RBF), HRBF (Hyper RBF), GPFN (Gaussian Potential Function Network) and SIGPI (Sigma-Pi Networks); as well as different adaptive learning algorithms: BP (Backpropagation), BPO (Backpropgation Online), DBD (Delta-Bar-Delta), SSAB (Super Self Adapting Backprop), RPROP (Resilient Propagation), MRPROP (modified RPROP), MSSAB (modified SSAB) [3, 9]. In computer simulations we have compared existing learning algorithms with new ones: MRPROP and MSSAB. The results indicate that the generalized RBF networks (ERBF, HRBF and SIGPI) with associated new learning algorithms converge faster and ensure better performance for general data-sets than standard models. In the simulations we used Matlab v. 4.2c and custom-made RBF simulator.

5. CONCLUSIONS

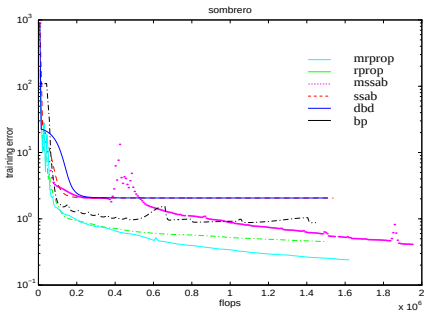
It has been shown that the mean integrated absolute error of recursive RBF nets converges to zero when the size of the network increases and parameters controlling the receptive field are simultaneously appropriately adjusted. Generalization of the results of this paper to general recursive RBF nets with positive definite matrices Σ is straightforward. More studies are needed on the analysis of recursive RBF nets with centers determined by clustering of the training sequence.

6. REFERENCES

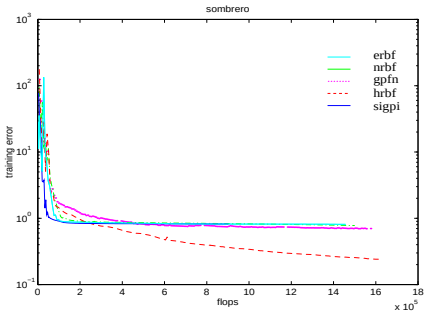
1. S. Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol.5, pp. 185-196, 1993.
2. D.S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks, *Complex Systems*, vol. 2, pp. 321-323, 1988.
3. A. Cichocki and R. Unbehauen, *Neural Networks for Optimization and Signal Processing*, Teubner Verlag-Wiley, Chichester, 1993.
4. F. Girosi and G. Anzellotti, "Rates of convergence for radial basis functions and neural networks," *Artificial Neural Networks for Speech and Vision*, R.J. Mammone, Ed., Chapman and Hall, London, 97-113, 1993.
5. F. Girosi, M. Jones and T. Poggio, "Regularization theory and neural networks architectures," *Neural Computation*, vol. 7, pp. 219-269, 1995.
6. W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990.
7. E.J. Hartman, J.D. Keeler and J.M. Kowalski, J.M., "Layered neural networks with Gaussian hidden units as universal approximations," *Neural Computation*, vol. 2, pp. 210-215, 1990.
8. A. Krzyżak and T. Linder, "Nonparametric estimation and classification using radial basis function nets and empirical risk minimization," *IEEE Trans. on Neural Networks*, vol. 7, pp. 475-487, 1996.
9. J. Mazurek, "Fast learning in RBF networks", Technical Report, 1997.
10. J. Moody and J. Darken, "Fast learning in networks of locally-tuned processing units," *Neural Computation*, vol. 1, pp. 281-294, 1989.
11. P. Niyogi and F. Girosi, "On the relationship between generalization error, hypothesis complexity, ans sample complexity for radial basis functions," *Neural Computation*, vol. 8, pp. 819-8442, 1996.
12. J. Park and I.W. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Computation*, vol. 5, pp. 305-316, 1993.
13. T. Poggio and F. Girosi, "A Theory of networks for approximation and learning," *Proc. IEEE*, vol. 78, pp. 1481-1497, 1990.
14. D.W. Scott, *Multivariate Density Estimation: Theory, Practice, Visualization* , Wiley, New York, 1992.
15. D.F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, pp. 109-118, 1990.
16. L. Xu, A. Krzyżak and E. Oja, "Rival Penalized Competitive Learning for Clustering Analysis, RBF net and Curve Detection," *IEEE Trans. on Neural Networks*, Vol.4, pp. 636-649, 1993.
17. L. Xu, A. Krzyżak and A.L. Yuille, "On radial basis function nets and kernel regression: approximation ability, convergence rate and receptive field size," *Neural Networks*, vol. 7, pp. 609-628, 1994.



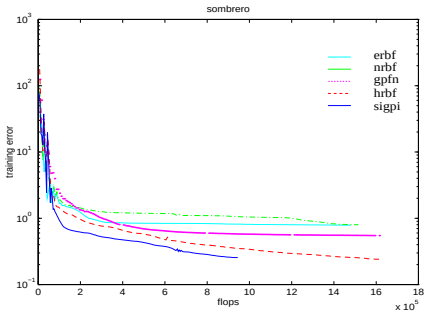
a) SRBF network



b) HRBF network

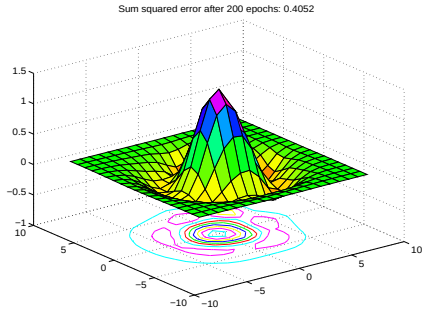


c) RPROP algorithm

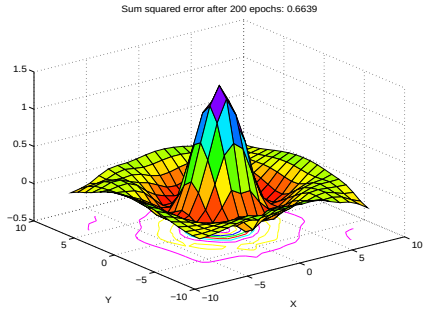


d) MRPROP algorithm

Figure 1. Training error versus number of flops for various training algorithms and network structures for sombrero function with 15 hidden neurons.

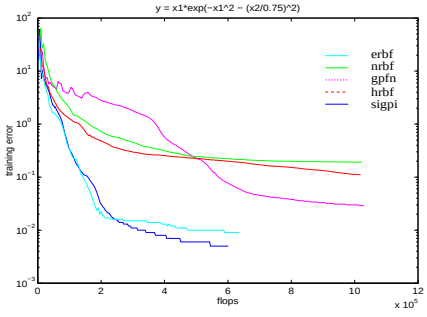


a) HRBF network

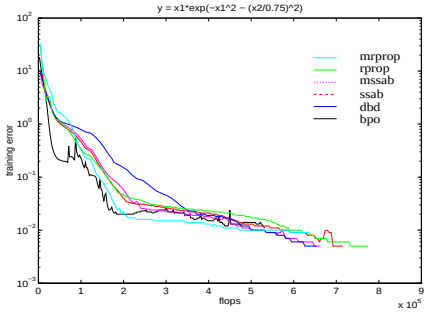


b) GPFN network

Figure 2. Approximation of sombrero function with mrprop algorithm and 15 hidden neurons after 200 epochs.



a) MRPROP algorithm



b) HRBF network

Figure 3. Approximation of wave function with 9 hidden neurons. Training error versus number of flops for various network structures and training algorithms.