

# Rates of Convergence for Fast Learning RBF Neural Networks

*Tassi di convergenza delle reti neurali RBF ad apprendimento rapido*

Diego Mancuso

Istituto di Statistica dell'Università Cattolica S.C., e-mail: [diego.mancuso@unicatt.it](mailto:diego.mancuso@unicatt.it)

**Riassunto.** Le reti neurali sono modelli di regressione non parametrica interessanti per la loro efficacia in presenza di un elevato numero di variabili esplicative. Tra i vari schemi di rete neurale quello di tipo RBF ad apprendimento rapido si caratterizza per un processo di stima estremamente semplice. In questo lavoro viene esplicitata la velocità di convergenza delle reti RBF ad apprendimento rapido mettendo in evidenza come, sotto tenui ipotesi di regolarità, questa non dipenda dal numero di regressori utilizzati.

**Keywords:** nonparametric regression, radial basis function neural networks.

## 1. Introduction

Let  $(X,Y)$  the support of  $d+1$  statistical variables ( $d$  of them called covariates and collected in the random vector  $\mathbf{x}$  and the last one in  $y$ ) and  $g(\mathbf{x},y)$  the absolutely continuous probability density function on  $(X,Y)$  given by the product of the marginal density of  $\mathbf{x}$  and the conditional density of  $y$ ,  $g(\mathbf{x},y) = g(\mathbf{x})g(y|\mathbf{x})$ . A nonparametric estimator of the regression function  $m(\mathbf{x}) = E_{y|\mathbf{x}}(y)$  is a  $R \times R^{n(d+1)} \rightarrow R$  function  $f_n(\mathbf{x},D)$  useful to recover  $m(\mathbf{x})$  by means of a random sample  $D$  of size  $n$  of the previous variables:  $D \equiv \{(\mathbf{x}_k, y_k) \in R^d \times R\}_{k=1}^n$ . In this paper we use standard matrix notation, thus  $E(.)$  and  $V(.)$  are the expectation and the variance operators,  $I$  is the identity matrix,  $\|\cdot\|$  is the Euclidean norm,  $tr(.)$  is the trace operator and so on. Moreover we assume the following hypotheses on  $D$ : a)  $\mathbf{X}$  is a bounded region of  $R^d$ ; b) the elements of  $D$ ,  $(\mathbf{x}_k, y_k)$  with  $k=1, \dots, n$ , are drawn independently from the support  $(X,Y)$ ; c)  $y_k = m(\mathbf{x}_k) + \xi_k$ ; d)  $\xi_k$  and  $\mathbf{x}_k$  are uncorrelated and  $\xi = [\xi_k]$  is a random variable with  $E(\xi) = \mathbf{0}$  and  $V(\xi) = \tau I$ ; e) the function  $m$  is continuous. Within this data framework, we are dealing with estimators of  $m$  which are linear models conditioning on the  $n$  observations  $\mathbf{x}_k$  of  $D$ :

$$f_n(\mathbf{x}, D | \mathbf{x}_k \in D) = \sum_{i=1}^n w_i(D) \phi_i(\mathbf{x}). \quad (1)$$

Several nonparametric estimators of the regression function belong to this class, but we focus only on one of these tools, the *fast learning radial basis function* (FL-RBF) neural networks. The FL-RBF neural nets present the following structure:

$$y_{RBF}(\mathbf{x}) = \sum_{i=1}^s w_i \phi(\|\mathbf{x} - \boldsymbol{\mu}_i\|)$$

where  $s < n$ ,  $\phi()$  is a radial basis function vanishing if  $\|\mathbf{x} - \mu_i\|$  tends to infinity and  $\mu_i \in \mathbb{R}^d$  are vectors. This kind of models stems from the particular approach to interpolation called *regularization* (Girogi and Poggio (1990)). Following this approach  $\phi()$  and the weights  $w_i$  are chosen by minimizing the functional

$$H[f_n] = \sum_{i=1}^n (y_i - f_n(\mathbf{x}_i))^2 + \alpha \|Pf_n\|^2$$

where  $P$  is a differential operator and  $\alpha$  is a positive smoothing parameter. The general solution of this variational problem leads to estimators with additive structure (1) and to choose  $\phi()$  as a Green's function of the differential operator  $P^*P$ ,  $P^*$  being the adjoint operator of  $P$ . This condition means that  $\phi()$  must satisfy  $P^*P\phi(\mathbf{x};\mathbf{t}) = \delta(\mathbf{x}-\mathbf{t})$  in the distribution sense, where  $\delta()$  is the *Dirac* function. In this paper we assume the function  $\phi()$  is a Green's function having no singularities in the origin for some differential operator  $P^*P$ . This is the case for the most popular choice of  $\phi()$ , the *gaussian* kernel  $\exp(-\|\mathbf{x} - \mu_i\|^2/2\sigma^2)$ . Moreover our point of view is asymptotic and thus the only meaningful way to set the vectors  $\mu_i$  in FL-RBF models is as a random sample of size  $s$  of the  $n$  observations  $\mathbf{x}_i$  in  $D$ . Defining  $\Phi = [\Phi_{ki}]$  the  $n \times s$  matrix which gathers the  $n$  values of the  $s$  radial basis functions along the observed  $\mathbf{x}_i$  of  $D$ , with

$$\Phi_{ki} = \phi_i(\mathbf{x}_k) = \phi(\|\mathbf{x}_k - \mu_i\|),$$

the weights  $\mathbf{w} = [w_i]$  are set as  $\mathbf{w} = (\Phi^T \Phi + \alpha I)^{-1} \Phi^T \mathbf{y}$ , where  $\mathbf{y} = [y_k]$ .

## 2. Conditional decomposition of the mean integrated squared error

An ordinary measure to compare different nonparametric regression techniques is the mean integrated squared error defined by  $Mise(f_n, m) = E_{\mathbf{x}}[E_{y|\mathbf{x}}(f_n(\mathbf{x}) - m(\mathbf{x}))^2]$ . Conditioning on the  $\mathbf{x}_k$  in  $D$  and by the well known bias-variance decomposition of the mean squared error, the *Mise* of a nonparametric estimator of  $m$  can be expressed as:

$$\begin{aligned} Mise(f_n, m) &= E_{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n} (E_{\mathbf{x}} E_{y|\mathbf{x}} [f_n(\mathbf{x}) - m(\mathbf{x})]^2 | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \\ &= E_{\mathbf{x}_1, \dots, \mathbf{x}_n} E_{\mathbf{x}} [E_{y|\mathbf{x}} f_n(\mathbf{x}) - m(\mathbf{x})]^2 | \mathbf{x}_1, \dots, \mathbf{x}_n + E_{\mathbf{x}_1, \dots, \mathbf{x}_n} E_{\mathbf{x}} [V_{y|\mathbf{x}} f_n(\mathbf{x}) | \mathbf{x}_1, \dots, \mathbf{x}_n]. \end{aligned} \quad (2)$$

Mancuso (2002) has already decomposed the conditional *Mise* of FL-RBF nets as

$$Mise(y_{RBF}, m) | \mathbf{x}_1, \dots, \mathbf{x}_n = E_{\mathbf{x}} [\gamma(\mathbf{x}, \omega_s)]^2 | \mathbf{x}_1, \dots, \mathbf{x}_n + O(s/n) \quad (3)$$

where  $\gamma(\mathbf{x}, \omega_s) = \sum_{j=1}^s \omega_j \phi_j(\mathbf{x}) - m(\mathbf{x})$  is the approximation error of the conditionally best FL-RBF network of  $s$  neurons identified by the vector of weights  $\omega_s = [\omega_j]$  and

$$E_{\mathbf{x}}[\phi_i(\mathbf{x})\gamma(\mathbf{x}, \omega_s)] = 0 \quad i = 1, \dots, s \quad (4)$$

In order to prove (3) we need only to notice that

$$\begin{aligned} E_{y|\mathbf{x}}(\mathbf{w}) &= \omega_s - \alpha(\Phi^T \Phi + \alpha I)^{-1} \omega_s - (\Phi^T \Phi + \alpha I)^{-1} [\Phi^T \gamma(\mathbf{x}, \omega_s)] \\ V_{y|\mathbf{x}}(\mathbf{w}) &= \tau \cdot [(\Phi^T \Phi + \alpha I)^{-1} - \alpha(\Phi^T \Phi + \alpha I)^{-2}] \end{aligned}$$

and from  $\lim n^{-1}(\Phi^T \Phi + \alpha I) = M$  a.s. (where  $M = [E_{\mathbf{x}}(\phi_i(\mathbf{x})\phi_j(\mathbf{x}))]$  is the stochastic moment matrix between radial basis functions) and  $\lim n^{-1} \Phi^T \gamma(\mathbf{x}, \omega_s) = E_{\mathbf{x}}[\phi(\mathbf{x})\gamma(\mathbf{x}, \omega_s)] = \mathbf{0}$  by (4), with  $\phi(\mathbf{x}) = [\phi_i(\mathbf{x})]$  and  $\gamma(\mathbf{x}, \omega_s) = [\gamma(\mathbf{x}_k, \omega_s)]$ , we obtain

$$E_{y|\mathbf{x}} f_n(\mathbf{x}) - m(\mathbf{x}) = \sum_{j=1}^s (E_{y|\mathbf{x}} w_j) \phi_j(\mathbf{x}) - m(\mathbf{x}) = \gamma(\mathbf{x}, \omega_s) + O(n^{-1}).$$

Now, the integrated conditional variance term in (2) can be computed recalling (1) as

$$\begin{aligned} E_{\mathbf{x}}[V_{y|\mathbf{x}} f_n(\mathbf{x})] | \mathbf{x}_1, \dots, \mathbf{x}_n &= E_{\mathbf{x}} \left[ \sum_{i=1}^s \sum_{j=1}^s \phi_i(\mathbf{x}) \text{Cov}(w_i, w_j) \phi_j(\mathbf{x}) \right] = \\ &= \sum_{j=1}^s \sum_{i=1}^s \text{Cov}(w_i, w_j) E_{\mathbf{x}}[\phi_i(\mathbf{x}) \phi_j(\mathbf{x})] = \sum_{i=1}^s \sum_{j=1}^s \text{Cov}(w_i, w_j) M_{ji} = \text{tr}[V_{y|\mathbf{x}}(\mathbf{w}) M], \end{aligned}$$

obtaining for FL-RBF nets

$$E_{\mathbf{x}}[V_{y|\mathbf{x}} f_n(\mathbf{x})] | \mathbf{x}_1, \dots, \mathbf{x}_n = \tau \cdot n^{-1} \text{tr}[(M)^{-1} M + O(n^{-1})] = \tau(s/n) + O(n^{-1}).$$

### 3. Rates of convergence for FL-RBF networks

In this final section we deduce the optimal rate of convergence for a FL-RBF neural net. First of all we introduce the space  $K$  of real functions with continuous derivatives of all orders and with bounded support. If  $\zeta \in K$ , starting from the identity  $\zeta(\mathbf{x}) = \int \zeta(\mathbf{t}) \delta(\mathbf{x} - \mathbf{t}) d\mathbf{t}$  and the assumed hypothesis that  $\phi()$  is a Green's function (therefore  $P^* P \phi(\mathbf{x}; \mathbf{t}) = \delta(\mathbf{x} - \mathbf{t})$ ), Girosi and Poggio (1990) show that

$$\zeta(\mathbf{x}) = \int_X \zeta(\mathbf{t}) (P^* P \phi)(\mathbf{x}; \mathbf{t}) d\mathbf{t} = \int_X \phi(\mathbf{x}; \mathbf{t}) (P^* P \zeta)(\mathbf{t}) d\mathbf{t} = \int_X \phi(\mathbf{x}; \mathbf{t}) \psi(\mathbf{t}) d\mathbf{t}$$

where  $\psi(\mathbf{x}) = (P^* P \zeta)(\mathbf{x})$  and  $P^*$  is the adjoint of the differential operator  $P$ . Now we can continue defining in (5) the random variable  $Z$ , where  $(\mathbf{x}_1, \dots, \mathbf{x}_s)$  is a sample on  $X$

$$Z(\mathbf{x}; \mathbf{x}_1, \dots, \mathbf{x}_s) = s^{-1} \sum_{j=1}^s \phi_j(\mathbf{x}) (\psi(\mathbf{x}_j) / g(\mathbf{x}_j)). \quad (5)$$

If the  $\mathbf{x}_j$  in  $(\mathbf{x}_1, \dots, \mathbf{x}_s)$  are drawn independently from  $g(\mathbf{x})$  we observe that

$$E_{\mathbf{x}_1, \dots, \mathbf{x}_s}(Z) = \zeta(\mathbf{x}) \quad \text{and} \quad V_{\mathbf{x}_1, \dots, \mathbf{x}_s}(Z) = s^{-1} V[\phi_j(\mathbf{x}) \psi(\mathbf{x}_j) / g(\mathbf{x}_j)] = O(s^{-1})$$

and applying (4), (5) and Cauchy-Schwartz inequality, we notice

$$\begin{aligned} E_{\mathbf{x}}[\gamma(\mathbf{x}, \omega_s)]^2 &= E_{\mathbf{x}} \left[ \gamma(\mathbf{x}, \omega_s) \left( \sum_{j=1}^s \phi_j(\mathbf{x}) \omega_j - Z \right) + \gamma(\mathbf{x}, \omega_s) (Z - m(\mathbf{x})) \right] = \\ &= \sum_{j=1}^s \left( \omega_j - \frac{1}{s} \frac{\psi(\mathbf{x}_j)}{g(\mathbf{x}_j)} \right) E_{\mathbf{x}}[\phi_j(\mathbf{x}) \gamma(\mathbf{x}, \omega_s)] + E_{\mathbf{x}}[\gamma(\mathbf{x}, \omega_s) (Z - m(\mathbf{x}))] = \\ &= E_{\mathbf{x}}[\gamma(\mathbf{x}, \omega_s) (Z - m(\mathbf{x}))] \leq \left( E_{\mathbf{x}}[\gamma(\mathbf{x}, \omega_s)]^2 \right)^{1/2} \left( E_{\mathbf{x}}[Z - m(\mathbf{x})]^2 \right)^{1/2}. \end{aligned}$$

The last inequality yields immediately  $E_{\mathbf{x}}[\gamma(\mathbf{x}, \omega_s)]^2 \leq E_{\mathbf{x}}[Z - m(\mathbf{x})]^2$  for any  $Z$  and then

$$\begin{aligned} E_{\mathbf{x}_1, \dots, \mathbf{x}_n} \left( E_{\mathbf{x}}[\gamma(\mathbf{x}, \omega_s)]^2 \mid \mathbf{x}_1, \dots, \mathbf{x}_n \right) &\leq E_{\mathbf{x}_1, \dots, \mathbf{x}_n} E_{\mathbf{x}}[(Z - \zeta(\mathbf{x}) + \zeta(\mathbf{x}) - m(\mathbf{x}))^2] = \\ &= E_{\mathbf{x}}[V(Z)] + E_{\mathbf{x}}[\zeta(\mathbf{x}) - m(\mathbf{x})]^2 = O(s^{-1}) + E_{\mathbf{x}}[\zeta(\mathbf{x}) - m(\mathbf{x})]^2. \end{aligned}$$

As Gel'fand and Shilov (1964) (p.141) show, if  $m$  is a continuous function there exists a sequence  $\{\zeta_s\}$  of functions in  $K$  converging uniformly to  $m$  in any bounded region. Now taking  $\{\zeta_s\}$  such that  $E_{\mathbf{x}}[\zeta_s(\mathbf{x}) - m(\mathbf{x})]^2 \leq O(s^{-1})$ , we get the generic rate of convergence for the unconditional *Mise* of a FL-RBF neural net:

$$Mise(y_{RBF}, m) = E_{\mathbf{x}_1, \dots, \mathbf{x}_n} \left( E_{\mathbf{x}}[\gamma(\mathbf{x}, \omega_s)]^2 \mid \mathbf{x}_1, \dots, \mathbf{x}_n \right) + O(s/n) \leq O(s^{-1}) + O(s/n).$$

This rate is not dependent on the number  $d$  of covariates and we can conclude that if we take  $s = O(n^{1/2})$  we obtain the optimal rate of convergence  $Mise(y_{RBF}, m) = O(n^{-1/2})$ .

## 4. Conclusion

Previous arguments prove the independence between  $L^2$  measures of fitting and the number  $d$  of covariates in FL-RBF interpolations. This result is coherent with other neural network schemes but surprising for the mild assumptions. It has relevant effects on variable selection techniques and on multivariate statistics. The obtained rate of convergence is better than *kernel* regression one when  $d$  is just moderately large.

## References

- Gel'fand I., Shilov G. (1964). *Generalized functions* vol.1, Academic Press, New York.  
 Girosi F., Poggio T. (1990), Networks and the best approximation property, *Biological Cybernetics*, 63, 169-176.  
 Mancuso D. (2002) Proprietà di base delle reti neurali RBF, in: *Studi in onore di Angelo Zanella*, Frosini B.V., Magagnoli U. & Boari G. (Eds), Vita e Pensiero, 439-454.