ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 303-308

Radial Basis Functions in the Fourier Domain

Mark J. L. Orr

Centre for Cognitive Science, University of Edinburgh, UK

Abstract. We demonstrate a relationship between the singular values of the design matrix and the discrete Fourier transform of the radial function for radial basis function networks. We then show how regularisation leads to high frequency filtering of the network output. In certain circumstances, this allows the network parameters to be chosen *a priori* to appropriately bias the learning process.

1 Introduction

Linear radial basis function networks modelling functions from $\mathbb{R}^n \rightarrow \mathbb{R}$ are characterised by a set of m centres $\{\mathbf{c}_j \in \mathbb{R}^n\}_{j=1}^m$ in a model of the form

$$f(\mathbf{x}) = \sum_{j=1}^{m} w_j \, h(\mathbf{x} - \mathbf{c}_j) \,, \tag{1}$$

where $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is a fixed symmetric function and the weights $\{w_j \in \mathbb{R}\}_{j=1}^m$ are adjustable parameters. The network learns from a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^p$ by minimising a cost function such as

$$C(\mathbf{w}) = \sum_{i=1}^{p} (y_i - f(\mathbf{x}_i))^2 + \lambda^2 \sum_{j=1}^{m} w_j^2 \,. \tag{2}$$

The first term encourages fidelity to the training set, the second discourages roughness in the output function and the two are balanced by the regularisation parameter λ. This is zero-order regularisation [5], also known as weight-decay [3] and ridge-regression [6]. Substituting (1) in (2), differentiating with respect to each weight and equating the results with zero locates the minimising weight vector, $\hat{\mathbf{w}} \in \mathbb{R}^m$, as the solution of the simultaneous equations

$$\sum_{i=1}^{p} f(\mathbf{x}_i) \, h(\mathbf{x}_i - \mathbf{c}_j) + \lambda^2 \, w_j = \sum_{i=1}^{p} y_i \, h(\mathbf{x}_i - \mathbf{c}_j) \,, \tag{3}$$

for $1 \leq j \leq m$. The solution is

$$\hat{\mathbf{w}} = \left(\mathbf{H}^\top \mathbf{H} + \lambda^2 \, \mathbf{I}\right)^{-1} \mathbf{H}^\top \mathbf{y} \,, \tag{4}$$

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 303-308

where $\mathbf{H} \in \mathbb{R}^{p \times m}$ is the design matrix, $H_{ij} = h(\mathbf{x}_i - \mathbf{c}_j)$, $\mathbf{I}$ the identity and $\mathbf{y} \in \mathbb{R}^p$ the vector of output training values. $\mathbf{H}$ can be replaced by its singular value decomposition,

$$\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^\top . \tag{5}$$

$\{\mathbf{u}_i \in \mathbb{R}^p\}_{i=1}^p$ (the columns of $\mathbf{U}$) and $\{\mathbf{v}_j \in \mathbb{R}^m\}_{j=1}^m$ (the columns of $\mathbf{V}$) are the orthonormal eigenvectors of $\mathbf{H}\mathbf{H}^\top$ and $\mathbf{H}^\top\mathbf{H}$, respectively. Since we assume $p \geq m$, there are m singular values, $\{s_j\}_{j=1}^m$, in descending order, lying on the diagonal of $\mathbf{S} \in \mathbb{R}^{p \times m}$. Substituting (5) in (4) and multiplying both sides by $\mathbf{V}^\top$ we obtain

$$\mathbf{v}_j^\top \hat{\mathbf{w}} = \frac{s_j}{\lambda^2 + s_j^2} \mathbf{u}_j^\top \mathbf{y} , \tag{6}$$

for $1 \leq j \leq m$. The network output over the training cases is $\mathbf{f} = \mathbf{H}\hat{\mathbf{w}}$. Multiplying both sides by $\mathbf{U}^\top$ and using (5) we get

$$\mathbf{u}_j^\top \mathbf{f} = \frac{s_j^2}{\lambda^2 + s_j^2} \mathbf{u}_j^\top \mathbf{y} , \tag{7}$$

for $1 \leq j \leq m$ and $\mathbf{u}_i^\top \mathbf{f} = 0$ for $m < i \leq p$.

2 Infinite Data and Centres

In the limit of infinite centres the model (1) becomes

$$f(\mathbf{x}) = \int_{\mathbb{R}^n} w(\mathbf{c})\, h(\mathbf{x} - \mathbf{c})\, d\mathbf{c} = (w * h)(\mathbf{x}) , \tag{8}$$

where '$*$' stands for convolution. The convolution is only possible due to the particular nature of radial functions and is not a feature of other types of bases such as logistic functions or polynomials. If the training data is also infinite then the m relations (3) become the single functional relation

$$((w * h) * h)(\mathbf{c}) + \lambda^2 w(\mathbf{c}) = (y * h)(\mathbf{c}) . \tag{9}$$

Taking Fourier transforms of both sides of (9) and employing the convolution theorem we can solve for the Fourier transform of the weights as

$$\hat{W}(\boldsymbol{\nu}) = \frac{H(\boldsymbol{\nu})}{\lambda^2 + H^2(\boldsymbol{\nu})} Y(\boldsymbol{\nu}) , \tag{10}$$

where $\boldsymbol{\nu}$ is a frequency vector in units of cycles per unit length (cpul). Then Fourier transforming (8) and using (10) gives

$$F(\boldsymbol{\nu}) = \frac{H^2(\boldsymbol{\nu})}{\lambda^2 + H^2(\boldsymbol{\nu})} Y(\boldsymbol{\nu}) . \tag{11}$$

Although equations (10) and (11) are idealisations obtained by replacing finite sums with continuous integrals, they bear an interesting resemblance to, respectively, equations (6) and (7). The resemblance suggests that the index j is related to frequency, and the singular values s_j are related to the Fourier transform $H(\boldsymbol{\nu})$. In the next section we investigate these relationships with the aid of a simulated learning problem.

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 303-308

3 Finite Data and Centres

Consider the following example: a scalar input ($n = 1$), $p = 100$ equally spaced training inputs $x_i \in [-10, 10]$, $m = 50$ equally spaced centres in the same range and a Gaussian radial function, $h(x) = \exp(-x^2/r^2)$, of width $r = 1.67$ (see section 4 for the choice of this value). Figure 1(a) shows a sample eigenvector, $\mathbf{v}_{10} \in \mathbb{R}^{50}$, corresponding to the 10th largest singular value, by plotting component values against centre positions.

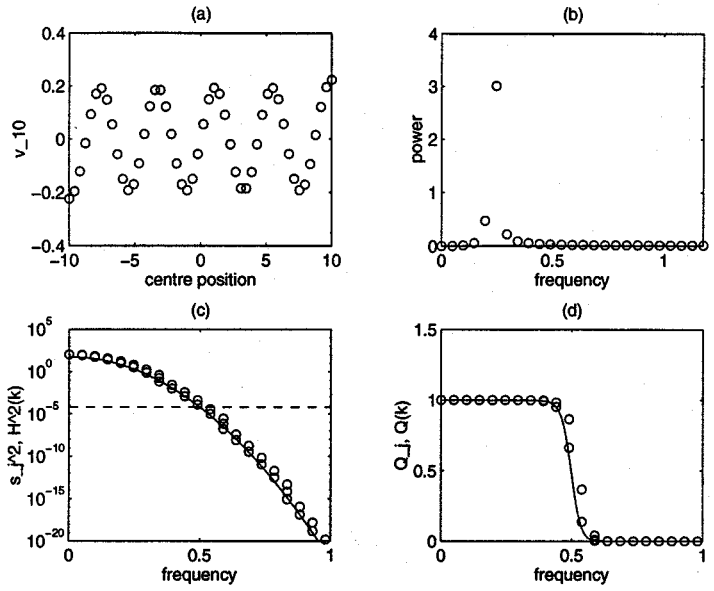

Figure 1: (a) The eigenvector of the 10th largest singular value. (b) Its discrete transform. (c) The dominant frequencies of the eigenvectors of the largest singular values (circles) compared with the spectrum of the radial function (solid line). The dashed line indicates the value of λ^2. (d) The high frequency filters Q_j (circles) and $Q(\nu_k)$ (solid line).

Note the sinusoidal shape with a dominant frequency of $\hat{\nu}_{10} = 0.25$ cpul, the frequency at which the discrete power spectrum attains a maximum (figure 1(b)). The other eigenvectors show similar structure with their dominant frequency, $\hat{\nu}_j$, depending on the size of their singular value, s_j. The (squared) singular values are plotted as small circles in figure 1(c) against dominant frequencies (up to the resolution of the discrete transforms—about 0.05 cpul). There is a close relation to the square of $\tilde{H}(\nu_k)$, the discrete Fourier transform of the radial function $h(x)$, shown by the solid curve. $\tilde{H}(\nu_k)$ was generated from samples with the same separation as the centres ($\Delta c = 20/(m-1)$) but taken over a wider range ($x \in [-40, 40]$) to obtain finer frequency resolution ($\Delta\nu \approx 0.01$ cpul) and avoid leakage effects [1]. The dashed line in figure 1(c) shows the size of the square of the regularisation parameter, $\lambda = 8 \times 10^{-3}$, and crosses the transform (solid curve) at about 0.5 cpul (see section 4).

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 303-308

The small circles in figure 1(d) plot $\hat{\nu}_j$, the dominant frequency of the j-th eigenvector $\mathbf{v}_j$, against

$$Q_j = \frac{s_j^2}{\lambda^2 + s_j^2}, \tag{12}$$

while the solid curve plots $\nu_k = k\,\Delta\nu$, the k-th frequency in the discrete transform, against

$$Q(\nu_k) = \frac{\tilde{H}^2(\nu_k)}{\lambda^2 + \tilde{H}^2(\nu_k)}. \tag{13}$$

The effect of the factors Q_j, which appear in equation (7), is to remove high frequency components from the output of the network. The critical frequency ν_c occurs where $s_j \approx \lambda$, the cross-over point in figure 1(c). Since s_j, as a function of frequency, is well approximated by $\tilde{H}(\nu_k)$ (figure 1(c)) and since the discrete transform $\tilde{H}(\nu_k)$ is a scaled (by $1/\Delta c$) version of the continuous transform $H(\nu_k)$ [1], it follows that the critical frequency can be estimated by solving $\lambda = H(\nu)/\Delta c$. In our example we used the Gaussian radial function $h(x) = \exp(-x^2/r^2)$ for which

$$H(\nu) = \sqrt{\pi}\, r\, \exp(-\pi^2 r^2\, \nu^2)\,.$$

The critical frequency is therefore

$$\nu_c \approx \frac{1}{\pi\, r}\left[\ln\left(\frac{\sqrt{\pi}\, r}{\lambda\, \Delta c}\right)\right]^{\frac{1}{2}}. \tag{14}$$

4 Choosing the Best Parameters

The RBF network as formulated in section 1 has three parameters: the size r of the radial function, the regularisation parameter λ and the number m (or separation Δc) of the centres (we assume the centres are arrayed over a fixed range). To show how the analysis above can be useful for the selection of these parameters, we extend the example in section 3 and consider a range of possible values for r and λ (keeping m fixed at 50 for simplicity).

Figure 2(a) shows a target function (solid curve) and $p = 100$ samples (dots) corrupted by white noise of size $\sigma = 0.25$. The circles in figure 2(b) show the discrete power spectrum of the sampled data. The white noise causes flattening of the spectrum at high frequencies at an expected level of $p\sigma^2 = 6.25$. Also shown in figure 2(b) are the discrete spectra of the target function (dashed curve) and the network output (solid curve) after learning from the sampled data. The network uses values for the parameters r and λ derived below.

Figures 2(c) and 2(d) are contour plots of critical frequency (14) and mean square error (MSE) as functions of r and λ. The MSE is between the target function and the

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 303-308

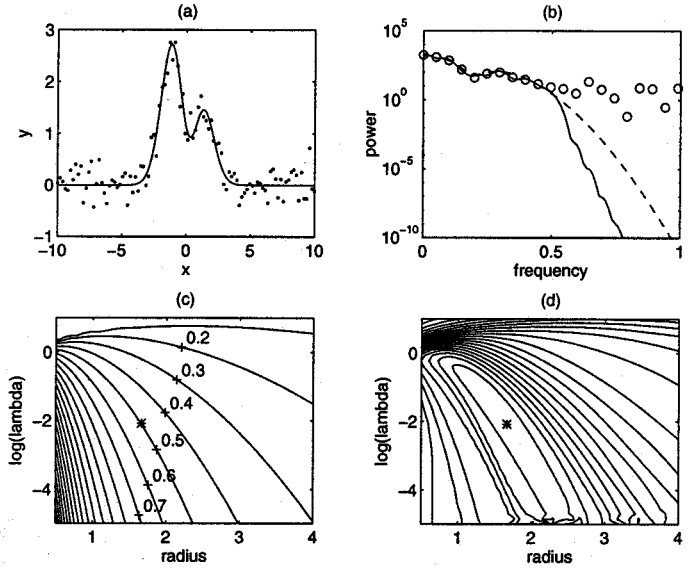

Figure 2: (a) The target function and a set of noisy samples. (b) The power spectra of the data (circles), the target function (dashed) and the network output (solid). (c) A contour plot of critical frequency (14) as a function of r and $\log(\lambda)$. (d) MSE (logarithmic contours) as a function of r and $\log(\lambda)$. A star—also shown in (c)—marks the position of the minimum.

network output over a large test set. The lowest MSE values in figure 2(d) are concentrated near the contour $\nu = 0.5$ cpul in figure 2(c), with the minimum value (marked with a star) near $r = 1.67$ and $\lambda = 0.008$. These were the parameter values used in the previous section and for the network whose output spectrum is shown in figure 2(b).

As r decreases along the $\nu_c = 0.5$ cpul contour line in figure 2(c) MSE in figure 2(d) gradually increases. This is due to widening of the radial function transform $H(\nu)$ and consequent raising of the low singular values such that the factors Q_j produce a less sharp cut-off and allow too much noise to filter through. At the other extreme, where r is increasing along the same contour, MSE suddenly becomes unstable because the transform becomes so narrow and the singular values (and λ) so small that the matrix inverse in equation (4) becomes numerically unstable.

5 Discussion

The analysis above suggests interpretations for the RBF parameters Δc, r and λ in terms of spectral analysis. The separation between centres, Δc, clearly plays the role of sampling rate and determines the maximum (Nyquist) frequency present in the unregularised output. The radial function width, r, sets the relative strength of the different frequencies

through their associated singular values. The radius and λ, the regularisation parameter, together determine the cut-off frequency. The best cut-off frequency, and consequently the best values for r and λ, depends on the lowest frequency at which the noise dominates the signal, which in the example is about 0.5 cpul (see figure 2(b)).

In real-world problems the input variable is often a vector ($n > 1$) and the training data rarely comes in a neat array. However, the conclusions remained unaltered after further simulations which we conducted with multiple dimensions and randomly chosen input points and centres. In particular, the best parameter values still corresponded to a cut-off frequency determined by the spectral characteristics of the signal and noise.

While an examination of the spectrum of the training data might prove an effective, if rather awkward, method of estimating the critical frequency (and hence the best network parameters) there are other much more convenient data-dependent ways of choosing good network parameters (e.g. cross-validation [4]). Consequently, we suggest the explicit relation between network parameters and critical frequency be used as a method of designing good bias into the learning process [2] from *a priori* knowledge of the critical frequency. Many practical problems will not have this kind of information available, but some will, particularly in signal processing and time series prediction.

For those problems which are able to take advantage of this and other kinds of *a priori* knowledge the burden on the data to convey all the information about the target function is reduced. This is particularly important when there is only sparse coverage of the input space, as in many of the most challenging problems in pattern recognition, and extrapolation rather than interpolation is required [2]. The work reported here can be viewed as a method of implementing a particular kind of knowledge (high frequency cut-off) in a particular kind of learner (regularised radial basis functions).

References

[1] E.O. Brigham. *The Fast Fourier Transform and its Applications.* Prentice Hall International, UK, 1988.

[2] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[3] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[4] M.J.L. Orr. Regularisation in the selection of radial basis function centres. *Neural Computation*, 7(3), 1995.

[5] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C.* Cambridge University Press, Cambridge, UK, second edition, 1992.

[6] J.O. Rawlings. *Applied Regression Analysis.* Wadsworth & Brooks/Cole, Pacific Grove, CA, 1988.