Learning and Generalisation in Radial Basis Function Networks

J.A.S. Freeman* D.Saad

Department of Physics
University of Edinburgh
Edinburgh EH9 3JZ
United Kingdom

This is the correct version to appear in *Neural Computation*.

Abstract

The two-layer Radial Basis Function network, with fixed centres of the basis functions, is analysed within a stochastic training paradigm. Various definitions of generalisation error are considered, and two such definitions are employed in deriving generic learning curves and generalisation properties, both with and without a weight decay term. The generalisation error is shown analytically to be related to the evidence and, via the evidence, to the prediction error and free energy. The generalisation behaviour is explored; the generic learning curve is found to be inversely proportional to the number of training pairs presented. Optimisation of training is considered by minimising the generalisation error with respect to the free parameters of the training algorithms. Finally, the effect of the joint activations between hidden-layer units is examined and shown to speed training.

*Email: jason@cns.ed.ac.uk

1 Introduction

Within the context of supervised learning in neural networks, one is primarily interested in minimising the average deviation of the actual network output from the desired output over the entire space of possible inputs. This quantity is not directly available within the paradigm of learning from a training set, and so is usually estimated with some approximation scheme, such as the mean sum-squared error on a set of test points which were not employed during training.

Generalisation error can be investigated analytically by making an assumption concerning the form of the function that generated the training set. One can then analyse properties of learning in the typical case, such as the decay rate of the generalisation error with number of training patterns and the optimal settings of parameters controlling the training algorithm. Several methods exist which facilitate such investigations, such as the VC and PAC frameworks (Vapnik and Chervonenkis, 1971, Haussler, 1994) and the statistical mechanics framework (see Watkin *et al.*, 1993, for a review). This paper utilises a Bayesian approach in which a probability distribution is constructed over the weight space of the network. Similar approaches can be found in Mackay, 1992 and Bruce and Saad, 1994.

To date, such analytic investigations of generalisation error have primarily focused on the one-layer perceptron, either in boolean or linear form, and on simple extensions of this, such as the committee machine (see, for instance, Schwarze 1993), as these architectures are analytically tractable unlike the general multi-layer perceptron. This paper calculates generalisation error for a more complicated network: the two-layer radial basis function network (RBF). The RBF is representationally powerful, being a universal approximator for continuous functions in the limit of an infinite number of hidden units (Hartman *et al.* 1990). It has been successfully employed in a number of applications, including chaotic time-series prediction (Casdagli 1989), speech recognition (Niranjan 1990) and data classification (Musavi *et al.* 1992).

Generalisation error for the RBF has been considered both analytically and empirically to some extent: Niyogi and Girosi, 1994, derive a bound under the assumption that the training algorithm always finds a globally optimal solution, but require only weak constraints on the function that generated the training set; they do not consider regularisation. This paper also contains an extensive bibliography pertaining to the topic of generalisation. Botros and Atkeson, 1991, compare the performance of various choices for the basis functions. Further afield, bounds have been derived for the case in which the hidden layer consists of units with sigmoidal transfer functions (Barron, 1993, 1994).

The typical training methodology employed for the RBF is to fix the parameters of the first layer utilising some algorithm to ensure that the positions of the training data in input space are adequately represented by the basis functions, and then either to solve a system of linear equations or use some training algorithm such as gradient descent to set the parameters of the second layer.

Training is computationally inexpensive as compared to multi-layer perceptrons.

This paper first presents a detailed specification of the RBF model to be analysed. Various definitions of generalisation error are then considered, and two such definitions selected for the analysis. The expressions for generalisation error are derived and linked to the evidence; finally the behaviour of the network is examined from several perspectives.

2 The Radial Basis Function Network

The RBF architecture consists of a two-layer fully-connected network (see figure 1), with an input layer which performs no computation. With no loss of generality, only a single output node is utilised in the analysis. Each hidden node is parameterised by two quantities: a centre $\mathbf{m}$ in input space, corresponding to the vector defined by the weights between the node and the input nodes, and a width σ_B^2. These parameters are assumed to be fixed by a suitable process, such as a clustering algorithm or maximising the likelihood of the parameters with respect to the training data. The activation function of the hidden nodes is radially symmetric in input space; the magnitude of the activation given a particular datapoint is a decreasing function of the distance between the input vector of the datapoint and the centre of the basis function. The distance metric employed is Euclidian. The role of the hidden units is to perform a non-linear transformation of the input space into the space of activations of the hidden units; it is this transformation that gives the RBF a much greater representational power than the linear perceptron. The output layer computes a linear combination of the activations of the basis functions, parameterised by the weights $\mathbf{w}$ between hidden and output layers.

Within this model, the basis functions will be taken as Gaussian; each hidden node will have a width σ_B^2 corresponding to the variance of the Gaussian. The overall function computed by the network is therefore:

$$f(\mathbf{x}, \mathbf{w}) = \sum_b w_b \exp\left(-\frac{\|\mathbf{x} - \mathbf{m}_b\|^2}{2\sigma_B^2}\right) \tag{1}$$

The training data D will be taken to consist of P input-output pairs indexed $1 \ldots p$: $(\mathbf{x_p}, y_p)$; the data will be assumed to be generated by a teacher RBF and corrupted under some noise process, with the input points being drawn from a symmetric Gaussian distribution of variance σ_X^2. The centres of the teacher will be taken to be identical to those of the student and to possess an identical width parameter σ_B^2. The fact that student and teacher centres are identical implies that the function to be learnt is exactly realizable. In the terminology of learning theory, this means that the approximation error is zero; the generalisation error is equivalent to the estimation error (see Niyogi and Girosi, 1994, for an overview).

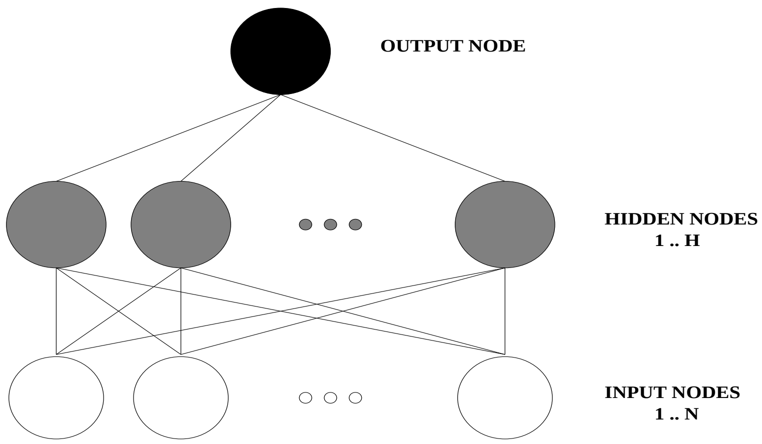


Figure 1: RBF network architecture

The training algorithm for the weights that impinge on the student output node will be considered stochastic in nature; this requires that an expression for the probability of a student weight vector given the training data and training algorithm parameters be defined. Modelling the noise process as zero-mean additive Gaussian noise leads to the following form for the probability of the dataset given the weights and training algorithm parameters:[1]

$$\mathcal{P}(D|\mathbf{w},\beta) = \frac{\exp(-\beta E_D)}{Z_D} \tag{2}$$

where $E_D = \frac{1}{2}\sum_p (y_p - f(\mathbf{x}_p, \mathbf{w}))^2$ is sum-squared training error and $Z_D = (2\pi/\beta)^{\frac{P}{2}}$.

This form resembles a Gibbs distribution over student space: it also corresponds to imposing the constraint that minimisation of the training error is equivalent to maximising the likelihood of the data (Levin *et al.* 1989). This distribution can be realized practically by employing the Langevin training algorithm, which is simply the gradient descent algorithm with an appropriate noise term added to the weights at each update (Rögnvaldsson, 1994). Furthermore, it has been shown that gradient descent, considered as a stochastic process due to random order of presentation of the training data, solves a Fokker-Planck equation for which the stationary distribution can be approximated by a Gibbs distribution (Radons *et al.* 1990).

To prevent over-dependence of the distribution of student weight vectors on the

[1] Note that, strictly, $\mathcal{P}(D|\mathbf{w},\gamma,\beta)$ should be written $\mathcal{P}((y_1,\ldots,y_P)|(\mathbf{x}_1,\ldots,\mathbf{x}_P),\mathbf{w},\gamma,\beta)$ as it is desired to predict the output terms from the input terms, rather than both jointly.

details of the noise, it is necessary to introduce a regularising factor, which can be viewed as a prior distribution over student space:

$$\mathcal{P}(\mathbf{w}|\gamma) = \frac{\exp(-\gamma E_W)}{Z_W} \tag{3}$$

where E_W is a penalty term based, for instance, on the magnitude of the student weight vector[2] and $Z_W = \int_W d\mathbf{w}\, \exp(-\gamma E_W)$.

Employing Bayes' theorem, one can derive an expression for the probability of a student weight vector given the training data and training algorithm parameters:

$$\begin{aligned} \mathcal{P}(\mathbf{w}|D,\gamma,\beta) &= \frac{\mathcal{P}(D|\mathbf{w},\beta)\,\mathcal{P}(\mathbf{w}|\gamma)}{\mathcal{P}(D|\gamma,\beta)} \\ &= \frac{\exp\left(-\beta E_D - \gamma E_W\right)}{Z} \end{aligned} \tag{4}$$

Here, $Z = \int d\mathbf{w}\, \exp(-\beta E_D - \gamma E_W)$ is the partition function over student space.

The quantity $\mathcal{P}(D|\gamma,\beta)$ has been termed the evidence for dataset D given the training algorithm parameters (MacKay 1992). It is proportional to the partition function, and thus closely related to the free energy, $F = -(1/\beta)\log Z$, an important quantity in the statistical mechanics framework (see, for instance, Hertz *et al.* 1989). It is of interest to relate analytically the evidence to generalisation error, as certain conjectures concerning this relation have been made on intuitive grounds (MacKay 1992).

3 Generalisation Error

There are several approaches that can be taken in defining generalisation error. The most prominent class of definitions focuses on the expectation of the difference between the desired network output and the actual output, as measured by some appropriate error measure, taken over the entire input space. The square of the difference between desired and actual output is the typical error measure employed, which for a particular student network gives:

$$E = \int_X d\mathbf{x}\,\mathcal{P}(\mathbf{x})\left(f(\mathbf{x},\mathbf{w}^0) - f(\mathbf{x},\mathbf{w})\right)^2 \tag{5}$$

where $\mathbf{w}^0$ is the weight vector of the teacher[3].

[2] Note that for the ubiquitous $E_W = \frac{1}{2}\|\mathbf{w}\|^2$ penalty term, $Z_W = (2\pi/\gamma)^{\frac{H}{2}}$.

[3] This definition is equivalent to the distance in the $L^2(\mathcal{P})$ norm between $f(\mathbf{x},\mathbf{w}^0)$ and $f(\mathbf{x},\mathbf{w})$, where $L^2(\mathcal{P})$ is the set of functions whose square is integrable with respect to the measure defined by P.

From a practical viewpoint, one only has access to the empirical risk, or test error, which is the mean-sum-squared error on a set of points not employed during training. This quantity is an approximation to the expected risk, defined as the expectation of $(y - f(\mathbf{x}, \mathbf{w}))^2$ with respect to the joint distribution $\mathcal{P}(\mathbf{x}, y)$. With an additive noise model, the expected risk simply decomposes to $E + \sigma_\eta^2$, where σ_η^2 is the variance of the noise. Some authors equate the expected risk with generalisation error by considering the squared difference between the *noisy* teacher and the student (see, for instance, Hansen 1993). A more detailed discussion of these quantities can be found in Niyogi and Girosi, 1994.

If a stochastic training algorithm is employed, such as the Langevin variant of gradient descent described previously, giving some probability distribution over weight space conditioned on the training data, there are two possibilities for the generalisation error. If, as is usually the case practically, the algorithm selects a single weight vector from the ensemble, a procedure which here will be termed Gibbs learning, then equation (5) becomes[4]:

$$E_G = \int_X d\mathbf{x}\, \mathcal{P}(\mathbf{x}) \int_W d\mathbf{w}\, \mathcal{P}(\mathbf{w}|D, \gamma, \beta)\, \left(f(\mathbf{x}, \mathbf{w}^0) - f(\mathbf{x}, \mathbf{w})\right)^2 \tag{6}$$

A second possibility arises from considering a Bayes-optimal approach. This requires one to take the expectation of the estimate of the network, which is impractical due to the computation involved, but can be approximated by performing a succession of training runs:

$$E_B = \int_X d\mathbf{x}\, \mathcal{P}(\mathbf{x}) \left(f(\mathbf{x}, \mathbf{w}^0) - \int_W d\mathbf{w}\, \mathcal{P}(\mathbf{w}|D, \gamma, \beta)\, f(\mathbf{x}, \mathbf{w})\right)^2 \tag{7}$$

These two quantities are related by:

$$\begin{aligned} E_G - E_B &= \int_W d\mathbf{w}\, \mathcal{P}(\mathbf{w}|D, \gamma, \beta)\, f(\mathbf{x}, \mathbf{w})^2 - \left(\int_W d\mathbf{w}\, \mathcal{P}(\mathbf{w}|D, \gamma, \beta)\, f(\mathbf{x}, \mathbf{w})\right)^2 \\ &= \mathrm{Var}[f(\mathbf{x}, \mathbf{w})] \end{aligned} \tag{8}$$

In order to investigate the generic performance of the architecture, it is desirable to eliminate the dependence of generalisation error on the particular data-set used. An average over possible data-sets, denoted by $\ll \cdots \gg$, will be utilised for this purpose. Thus, with additive Gaussian noise η on the data, one obtains:

[4]It is worth noting that by taking $\frac{\beta}{\gamma} \to \infty$, the distribution of student weight vectors becomes a delta function centred on the weight vector that minimises the empirical risk. This situation is commonly considered in the computational learning theory literature, but is unrealistic for neural networks, where often in practice only locally optimal solutions are found.

$$\ll E \gg = \int_{X^P} d^P\mathbf{x}\, \mathcal{P}(\mathbf{x}_p) \int_{\eta^P} d^P\eta\, \mathcal{P}(\eta_p) \int_X d\mathbf{x}\, \mathcal{P}(\mathbf{x}) \, \left(f(\mathbf{x}, \mathbf{w}^0) - f(\mathbf{x}, \mathbf{w})\right)^2 \quad (9)$$

An alternative measure of generalisation performance is a quantity known as prediction error (Levin *et al.* 1989), $E_P = -\log \mathcal{P}(y|\mathbf{x}, D)$, which is derived from the probability of the network correctly predicting a data-point drawn from a known probability distribution. Prediction error is closely linked to both the free energy F and the evidence.

4 Calculation of Generalisation Error

The calculation of generalisation error will focus on both E_G and E_B; a link to prediction error is developed via an analytic relation between E_G and the evidence.

Recalling that the teacher centres are equal in number and position to those of the student and signifying the difference between student and teacher weight vectors, $\mathbf{w} - \mathbf{w}^0$, by $\mathbf{w}^*$, the definition of E_G becomes:

$$\ll E_G \gg = \ll \int_W \mathcal{P}(\mathbf{w}|D, \gamma, \beta) \int_X \mathcal{P}(\mathbf{x}) \left[\sum_b w_b^* \exp\left(-\frac{\|\mathbf{x} - \mathbf{m}_b\|^2}{2\sigma_B^2} \right) \right]^2 d\mathbf{x}\, d\mathbf{w} \gg \quad (10)$$

Since the input vectors are drawn from a symmetric Gaussian distribution with mean $\mathbf{0}$, variance σ_X^2, on performing the integral over input space one obtains:

$$\ll E_G \gg = a \ll \int_W \mathcal{P}(\mathbf{w}|D, \gamma, \beta) \sum_{bc} w_b^* w_c^* G_{bc}\, d\mathbf{w} \gg \quad (11)$$

where:

$$\begin{aligned} a &= \left(\frac{\sigma_B^2}{\sigma_B^2 + 2\sigma_X^2} \right)^{\frac{N}{2}} \\ G_{bc} &= \exp\left(\frac{\|\mathbf{m}_b + \mathbf{m}_c\|^2 \left(\frac{\sigma_X^2}{\sigma_B^2 + 2\sigma_X^2} \right) - \|\mathbf{m}_b\|^2 - \|\mathbf{m}_c\|^2}{2\sigma_B^2} \right) \end{aligned}$$

with $\mathbf{m}_b$, $\mathbf{m}_c$ referencing the centres.

Employing the definition of $\mathcal{P}(\mathbf{w}|D, \gamma, \beta)$ as in equation (4), taking E_D as sum-squared training error with η_p as the noise on training example p, and defining

$E_W = \frac{1}{2}\|\mathbf{w}\|^2$ as the prior over weight space allows equation (4) to be re-written as:

$$\mathcal{P}(\mathbf{w}|D,\gamma,\beta) = \frac{\exp(-\frac{P}{2}\mathbf{w}^{*T}\Lambda^{-1}\mathbf{w}^* - \mathbf{w}^*\rho - \frac{\beta}{2}\sum_p \eta_p^2 - \frac{\gamma}{2}\|\mathbf{w}^0\|^2)}{Z} \tag{12}$$

where:

$$\begin{aligned} (\Lambda^{-1})_{bc} &= \frac{1}{P}\left(\gamma\delta_{bc} + \beta\sum_p q_b(p)q_c(p)\right) \\ \rho_b &= \gamma w_b^0 + \beta\sum_p \eta_p q_b(p) \\ q_b(p) &= \exp\left(-\frac{\|\mathbf{x}_p - \mathbf{m}_b\|^2}{2\sigma_B^2}\right) \end{aligned}$$

Now by taking the derivative of the numerator of equation (12) with respect to the elements of the matrix Λ^{-1}, $\ll E_G \gg$ becomes:

$$\ll E_G \gg = -\frac{2a}{P}\sum_{bc} G_{bc} \ll \frac{1}{Z}\frac{\partial}{\partial\Lambda_{bc}^{-1}}\left[\int_W \exp(-\gamma E_W - \beta E_D)\right] \gg \tag{13}$$

Recalling that the evidence is proportional to the partition function Z, one can immediately relate the evidence to the generalisation error:

$$\ll E_G \gg = -\frac{2a}{P}\sum_{bc} G_{bc} \ll \frac{\partial}{\partial\Lambda_{bc}^{-1}}[\log\mathcal{P}(D|\gamma,\beta)] \gg \tag{14}$$

At this point it is also possible to relate generalisation error to prediction error. It is relatively simple to derive the relationship between prediction error and evidence:

$$\log\mathcal{P}(y|\mathbf{x},D,\gamma,\beta) = \log\mathcal{P}(D,y|\mathbf{x},\gamma,\beta) - \log\mathcal{P}(D|\gamma,\beta) \tag{15}$$

Employing this relationship in equation (14), one arrives at:

$$\ll E_G \gg = -\frac{2a}{P}\sum_{bc} G_{bc} \ll \frac{\partial}{\partial\Lambda_{bc}^{-1}}[\log\mathcal{P}(D,y|\mathbf{x},\gamma,\beta)] - \frac{\partial}{\partial\Lambda_{bc}^{-1}}[\log\mathcal{P}(y|\mathbf{x},D,\gamma,\beta)] \gg \tag{16}$$

Returning to the derivation of E_G, calculating the evidence and performing the partial derivatives of equation (14) leads to:

$$\ll E_G \gg = \frac{a}{P} \ll \operatorname{tr} G\Lambda + \frac{1}{P}\rho\Lambda G\Lambda\rho \gg \tag{17}$$

It remains to consider the average $\ll \cdots \gg$ over datasets and the Gaussian noise on the datasets. Performing the noise average, recalling that only ρ contains noise terms:

$$
\int d^P\eta\, \mathcal{P}(\eta_p)\rho\Lambda G\Lambda\rho = \sum_{de} \Lambda G\Lambda_{de} \left(\gamma^2 w_d^0 w_e^0 + \beta^2\sigma_\eta^2 \sum_p q_d(p)q_e(p)\right) \tag{18}
$$

To progress further and perform the dataset average, it is necessary to know the form of Λ. To this end, it will be assumed that Λ^{-1} is of the form:

$$
\begin{bmatrix} \theta & \tilde{\theta} & \cdots & \tilde{\theta} \\ \tilde{\theta} & \theta & \cdots & \tilde{\theta} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\theta} & \tilde{\theta} & \cdots & \theta \end{bmatrix} \tag{19}
$$

That is, all diagonal entries are equal to θ, and all off-diagonal entries are equal to $\tilde{\theta}$.

This induces Λ to take on the form:

$$
\begin{bmatrix} \psi & \tilde{\psi} & \cdots & \tilde{\psi} \\ \tilde{\psi} & \psi & \cdots & \tilde{\psi} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\psi} & \tilde{\psi} & \cdots & \psi \end{bmatrix} \tag{20}
$$

where:

$$
\begin{aligned}
\psi &= \frac{1+\phi}{\theta-\tilde{\theta}} \\
\tilde{\psi} &= \frac{\phi}{(\theta-\tilde{\theta})} \\
\phi &= -\frac{\tilde{\theta}}{\theta+\tilde{\theta}(H-1)}
\end{aligned}
$$

The implications of this assumption for the RBF model are twofold: firstly, the equality of diagonal entries corresponds to all the centres receiving an equal amount of activation via the training data[5]. For the particular case of a symmetric input distribution centred at the origin of input space, this assumption

[5]The authors thank an anonymous referee for pointing out that a common procedure for

breaks down only for the case in which the centres are dissimilar in distance from the origin *and* the variance of the input distribution is not of sufficient magnitude for the distribution to be approximately uniform in the regions covered by the basis functions. Secondly, the equality of off-diagonal entries requires each pair of basis functions to receive a similar joint activation via the training data. This assumption is satisfied except for the case in which the centres are not approximately equidistant from each other *and* the spread of the basis functions is not sufficient to allow considerable overlap between each pair of receptive fields to occur.

Unfortunately, this selection of form for Λ^{-1} is not sufficient to allow the dataset average to be carried out, as the $\mathbf{x}_p$'s do not separate into independent factors. One can approximate Λ^{-1} as:

$$\begin{aligned}\Lambda_{bc}^{-1} &= \frac{\gamma\delta_{bc}}{P} + \frac{\beta}{P}\sum_p q_b(p)q_c(p) \\ &\simeq \frac{\gamma\delta_{bc}}{P} + \beta\left\langle \frac{1}{P}\sum_p q_b(p)q_c(p)\right\rangle_{X^P} \\ &= \frac{\gamma\delta_{bc}}{P} + \beta a\, G_{bc}\end{aligned} \tag{21}$$

where $\langle\cdots\rangle_{X^P}$ denotes an average over datasets. Utilising the central limit theorem, the neglected variance in the distribution of $\frac{1}{P}\sum_p q_b(p)q_c(p)$ decreases as $\frac{1}{P}$. Note that this implies that the calculation of generalisation error holds strictly only in the asymptotic regime of large P, but it will be shown via simulations that the results are a good approximation for non-asymptotic P.

The integral over datasets can now be performed as a straightforward Gaussian, yielding the final expression for generalisation error:

$$\ll E_G \gg = \frac{a}{P}\left(\operatorname{tr} G\Lambda + \frac{1}{P}\operatorname{tr}\Lambda G\Lambda\Upsilon\right) \tag{22}$$

where, for notational convenience, the matrix defined by $\Upsilon_{bc} = \gamma^2 w_b^0 w_c^0 + \beta^2\sigma_\eta^2 PaG_{bc}$ has been introduced.

From this, via equation 3.4, one can calculate $\ll E_B \gg$:

$$\ll E_B \gg = \frac{a}{P^2}(\operatorname{tr}\Lambda G\Lambda\Upsilon) \tag{23}$$

selecting basis function parameters is to maximise the likelihood of the inputs of the training data under a mixture model given by a linear combination of the basis functions; constraining the priors of the mixture model to be equal encourages this property of equal activation to be satisfied.

To examine the validity of the assumptions for Λ^{-1}, simulations were conducted in which the empirical value of E_G was calculated via equation (18) by generating random training data and numerically evaluating Λ. The simulations were carried out for three scenarios: firstly, the case in which the conditions for the assumption of form of Λ^{-1} were exactly satisfied; secondly, for certain basis functions receiving an impoverished supply of training data, thus violating the equality of diagonal entries; finally, for the interactions between different pairs of basis functions being unequal, which violates the equality of off-diagonal entries[6].

Comparisons of the mean values of E_G found by simulation, E_G^{SIM}, with those found analytically via equation (22) are shown in figure 2. Note that the variances of the simulation distributions quickly become negligible.

When the assumptions are satisfied, E_G^{SIM} rapidly converges to E_G. Violation of the assumption of diagonal equality gives rise to a systematic error, while violation of the off-diagonal assumption causes the convergence to slow, but introduces negligible systematic error. This lack of significant effect is explicable by an examination of the definition of G: the result of introducing differing interactions between the basis functions is simply to vary $\|\mathbf{m}_b + \mathbf{m}_c\|$; the effect of this will always be overwhelmed by that of other terms, particularly if the ratio of σ_B^2 to σ_x^2 is large. It can be concluded, therefore, that the calculation of generalisation error is invalid only for the cases in which P is near to 0 or in which the basis functions receive significantly different levels of activation via the training data.

5 Analysis of Generalisation Error

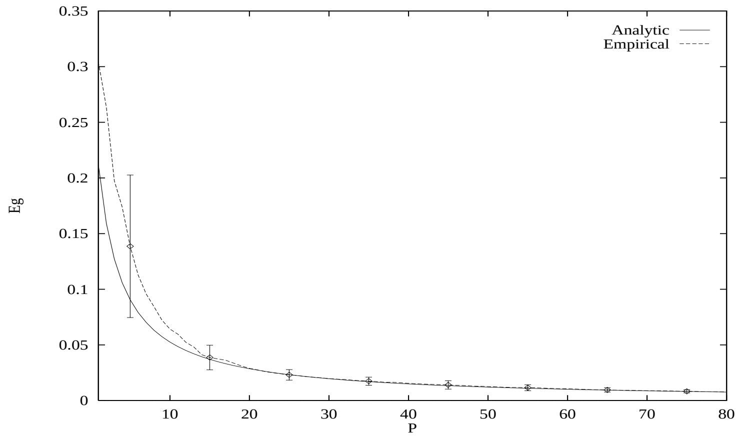
The equations derived for E_G and E_B do not admit to a straightforward intuitive understanding of the effect of varying parameters such as the number of training patterns, noise level and training parameters γ and β.

In order to promote such an understanding, the behaviour of the expressions for generalisation error will initially be examined under simplifying limiting conditions.

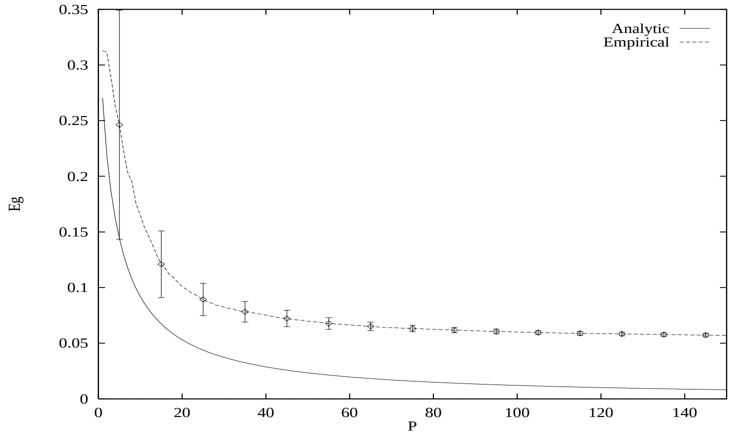
5.1 Noiseless Training Data

Taking the $\sigma_\eta^2 \to 0$ limit while treating β as a free parameter leads to the conclusion that, for both E_G and E_B, optimal training occurs when $\beta \to \infty$

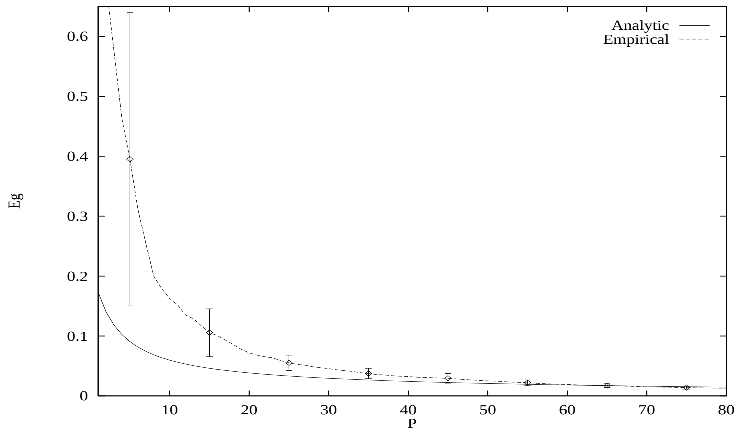
[6] Each simulation was run 50 times with the following parameter settings (denoting the angle between $\mathbf{m}_b$ and $\mathbf{m}_c$ as $\Theta_{b,c}$):
Common to all simulations: $N = 3, H = 4, \sigma_\eta^2 = 1, \beta = 0.5, \gamma = 1, \sigma_X^2 = 2, \sigma_B^2 = 1$
Assumptions satisfied: $\forall_b : \|\mathbf{m}_b\| = 1, \forall_{b,c:b\neq c} : \Theta_{b,c} = 2\pi/3$
Diagonal violation: $\|\mathbf{m}_1\| = \|\mathbf{m}_2\| = 1$, $\|\mathbf{m}_3\| = \|\mathbf{m}_4\| = 4$, $\forall_{b,c:b\neq c} : \Theta_{b,c} = 2\pi/3$
Off-diagonal violation: $\forall_b : \|\mathbf{m}_b\| = 1$, $\Theta_{1,2} = \Theta_{3,4} = \pi/6$, $\Theta_{1,4} = \Theta_{2,3} = \pi$



(a) Assumptions satisfied



(b) Violation of Diagonal Assumption



(c) Violation of Off-Diagonal Assumption

Figure 2: Analytic E_G (unbroken line) versus mean of E_G^{SIM} (dashed line) examining the validity of the assumption of form for Λ^{-1} under various distributions of the centres of the basis functions. The error bars are plotted at 1 standard deviation of the simulation mean.

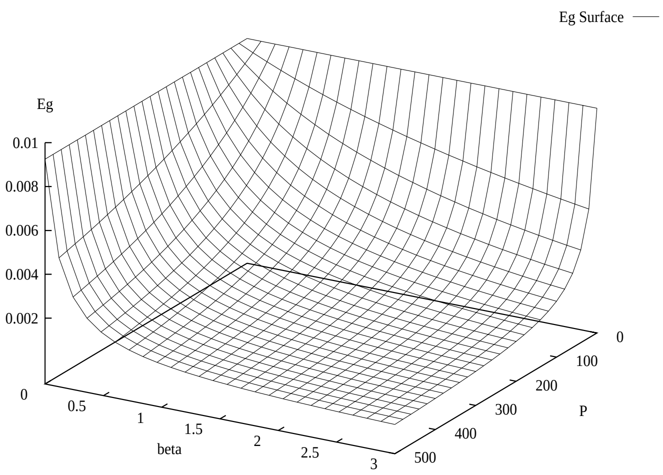


Figure 3: E_G as a function of number of examples P and error sensitivity β for $\sigma_\eta^2 \to 0$

(see figure 3). This is intuitively plausible; if the training data is not noisy then no training error should be tolerated, so forcing the distribution over student space to become a delta function centred on the value of $\mathbf{w}$ that sets the error to zero is reasonable. Note that in the $\beta \to \infty$ limit, the prior on student space becomes irrelevant.

5.2 No Weight Decay: the $\gamma \to 0$ limit

Considering the $\gamma \to 0$ limit allows one to analyse the dependence of E_G and E_B on the number of training examples, P. The assumption of the diagonal versus off-diagonal form for Λ^{-1} induces a similar form on the matrix G; referencing the diagonal and off-diagonal elements of G by G_D and G_O respectively and defining the matrix Ω by $\Omega_{bc} = \delta_{bc} + \lim_{\gamma\to 0}\phi$, which is both P and β independent, one obtains:

$$\ll E_G \gg \stackrel{\gamma\to 0}{=} \frac{\operatorname{tr} G\Omega}{\beta P(G_D - G_O)} + \frac{\sigma_\eta^2 \operatorname{tr} \Omega G \Omega G}{P(G_D - G_O)^2} \tag{24}$$

and

$$\ll E_B \gg \stackrel{\gamma \to 0}{=} \frac{\sigma_\eta^2 \operatorname{tr} \Omega G \Omega G}{P(G_D - G_O)^2} \tag{25}$$

It is apparent that both E_G and E_B are inversely proportional to the number of training examples. This result is somewhat similar to that found for the linear perceptron in this limit, whereby E_G and E_B are inversely proportional to $P - N - 1$ (Hansen 1993, Bruce and Saad 1994).

In addition, the $\gamma \to 0$ limit brings to light an interesting difference between E_G and E_B. Examining E_B, it is apparent that β plays no role; the expression is independent of the error sensitivity. This result is in contrast to that for E_G, in which the first term is minimised by taking $\beta \to \infty$. This hints that, in the Bayes generaliser, it is only the *ratio* of γ to β that is important, as is the case for the linear perceptron (Bruce and Saad 1994), while the Gibbs generaliser is dependent on both β and γ separately. This discrepancy is explicated by recalling equation (8); E_G consists of a variance term, minimised by taking $\beta \to \infty$, and a term identical to E_B.

Both E_G and E_B are independent of N, the dimensionality of input space, in this limit.

5.3 The General Case: Noise and Weight Decay

To gain some understanding of the variation of E_G and E_B with P, γ and β in the general case, consider figures 4, 5, 6 and 7.

Examining first figure 4, in which E_B is plotted against P and β for a constant value of γ, it is apparent that there is a minimum in the generalisation error surface at a constant value of β. When γ is set to its optimal value, the value of β at the minimium can be shown empirically to be inversely proportional to the variance of the noise, σ_η^2. Similarly, plotting E_B against P and γ (figure 5) demonstrates a minimum in the generalisation error surface at a constant value of γ. This minimum, for β set to an optimal value, is a function of both $\|\mathbf{w}^0\|^2$ and $\sum_{bc} \mathbf{w}_b^0 \mathbf{w}_c^0$.

An entirely different pattern of results emerges for E_G. Considering figure 6, the optimal value of β rapidly becomes infinite as P increases. This discrepency is due to the fact that the Gibbs generaliser requires the selection of a single weight vector from the ensemble of students, so it is advantageous to penalise any training error maximally once a reasonable amount of training data is available. The Bayes generaliser, on the other hand, employs a weighted average of students in order to make a prediction; noise on the training data output values can to some extent be compensated for by this average, and so it is not desirable to force the ensemble to become a delta function. Focussing on E_G as a function of P and γ (figure 7), an analogous result is apparent: the optimal value of γ is initially infinite, but as $P \to \infty$, the optimal value of γ tends to an expression similar in dependence to that for E_B.

5.4 Analytic Determination of Optimal Parameters

It is not possible to find closed-form analytic expressions for the optimal settings of β and γ for either E_G or E_B generally, but for the case in which there is no interaction between the basis functions, as may occur when the variance of the input distribution is large compared to the width of the basis functions, such expressions can be obtained; these can then be elaborated upon to some extent in order to suggest the form of the actual dependencies of β_{opt} and γ_{opt}.

For the Bayes-optimal generaliser, by minimising $\ll E_B \gg$ with respect to the training parameters, the optimal settings were determined to be:

$$\beta_{opt} = \frac{\gamma \|\mathbf{w}^0\|^2}{H\sigma_\eta^2} \tag{26}$$

and:

$$\gamma_{opt} = \frac{H\sigma_\eta^2 \beta}{\|\mathbf{w}^0\|^2} \tag{27}$$

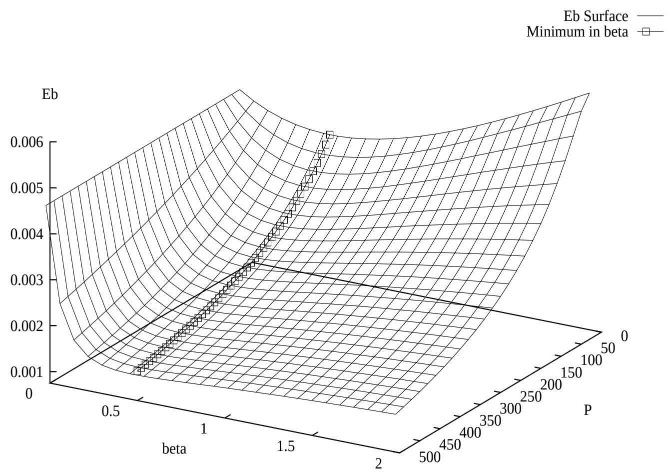


Figure 4: Generalisation error E_B as a function of number of examples P and error sensitivity β. The minimum in E_B with respect to β is independent of P.

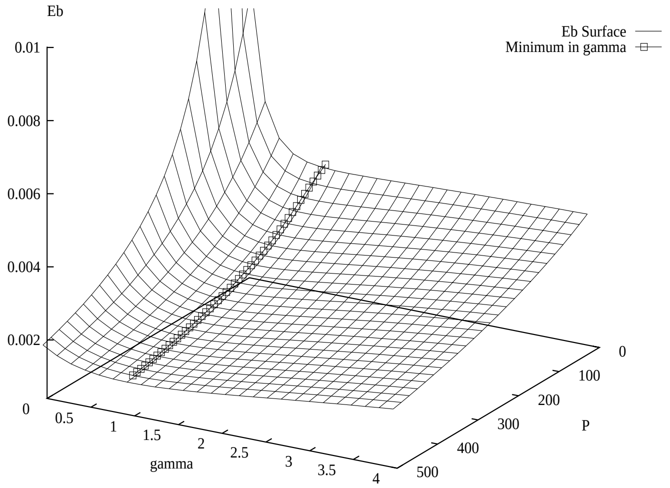


Figure 5: Generalisation error E_B as a function of number of examples P and weight decay parameter γ. The minimum in E_B with respect to γ is independent of P.

The form of equations (26) and (27) proves that only the *ratio* of γ to β, $\frac{2H\sigma_\eta^2}{\|\mathbf{w}^0\|^2}$, determines whether the parameter settings are optimal.

For the Gibbs generaliser the expressions for optimal parameters are a little more complicated:

$$\beta_{opt} = \frac{\gamma(2\gamma\|\mathbf{w}^0\|^2 + H)}{H(2\gamma\sigma_\eta^2 - aG_D P)} \tag{28}$$

$$\gamma_{opt} = \frac{aG_D HP\beta(2\beta\sigma_\eta^2 + 1)}{2\|\mathbf{w}^0\|^2\beta a G_D P - H} \tag{29}$$

Under this assumption of no interactions between the basis functions, the results for optimal parameters closely resemble those found for the perceptron (Bruce and Saad 1994), an architecture which can also be viewed as having no interactions between units of the layer immediately preceeding the output layer.

Allowing terms linear in the interaction parameter, G_O, leads to optimal parameters which have an additional dependence on the cross-correlation of the

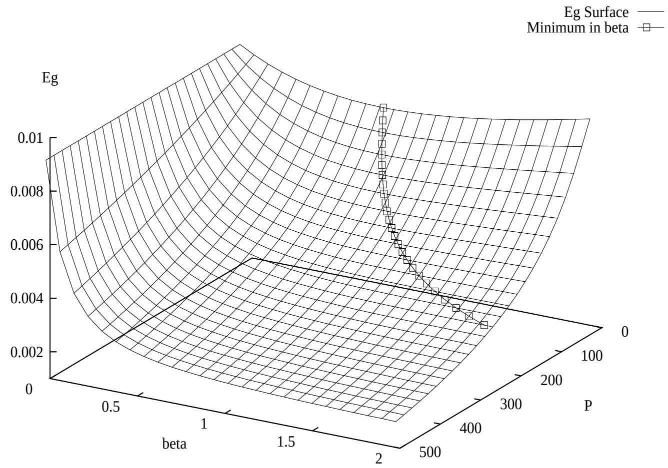


Figure 6: Generalisation error E_G as a function of number of examples P and error sensitivity β. At the minimum in E_G with respect to β, $\beta \to \infty$ as $P \to \infty$.

teacher RBF weight vector, $\sum_{bc} w_b^0 w_c^0$. For instance, the optimal ratio of γ_{opt} to β_{opt} for E_B becomes (with G_D small):

$$\frac{\gamma_{opt}}{\beta_{opt}} = \frac{2\beta\sigma_\eta^2 H G_D^2}{(G_D - G_O)G_O \sum_{bc} w_b^0 w_c^0 + (G_D - G_O)^2 \|\mathbf{w}^0\|^2} \tag{30}$$

The effect of admitting all terms in G_O for E_B can only be examined empirically. As in the $G_O = 0$ case, β_{opt} was found to be linearly dependent on γ, and vice versa, with the gradient of the γ_{opt} versus β dependence being the reciprocal of that for β_{opt} versus γ. This form of relationship implies that E_B can still be minimised by finding the correct ratio of γ to β; it is unnecessary to find absolute values for these quantities. Thus, the optimal values define a straight line in training parameter space.

In the case of E_B, the dependence of γ_{opt} and β_{opt} on the noise variance σ_η^2 can also be found; again, as in the $G_O = 0$ case, γ_{opt} is proportional to σ_η^2 while β_{opt} is inversely proportional to σ_η^2.

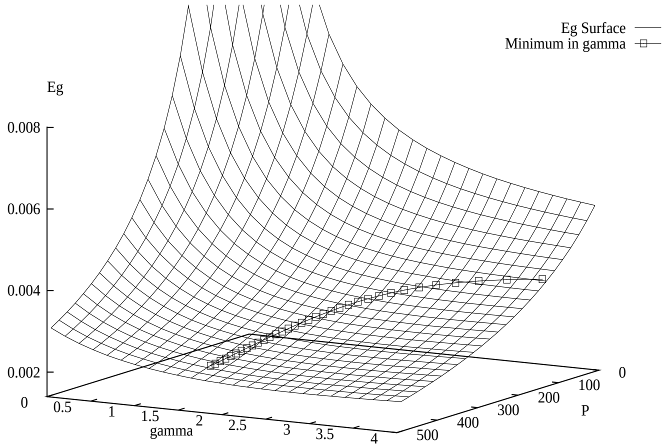


Figure 7: Generalisation error E_G as a function of number of examples P and weight decay parameter γ. As $P \to \infty$, the value of γ at the minimum in E_G with respect to γ becomes constant.

5.5 Interactions Between Hidden-Layer Units

The effect of joint activations between hidden-layer units, whereby a single training pair simultaneously contributes to the activation of every hidden-layer unit, is to reduce the number of training patterns required to achieve a certain level of generalisation error as compared to a network in which there are no such interactions. Consider figure 8, in which E_G is plotted for an RBF network with highly-overlapping hidden units and for a network with small overlap: the generalisation error for given P is considerably lower for the highly-overlapping version. This phenomenon is due to the fact that high overlaps allow every hidden unit to learn from every training pair, while small overlaps prevent some units from benefitting from certain training pairs.

6 Conclusion

Learning and generalisation in Radial Basis Function networks has been investigated via the assumption of a form for the function which generated the training data. By fixing the centres of the student basis functions to be equal to those of the teacher and employing a stochastic training paradigm for the output node

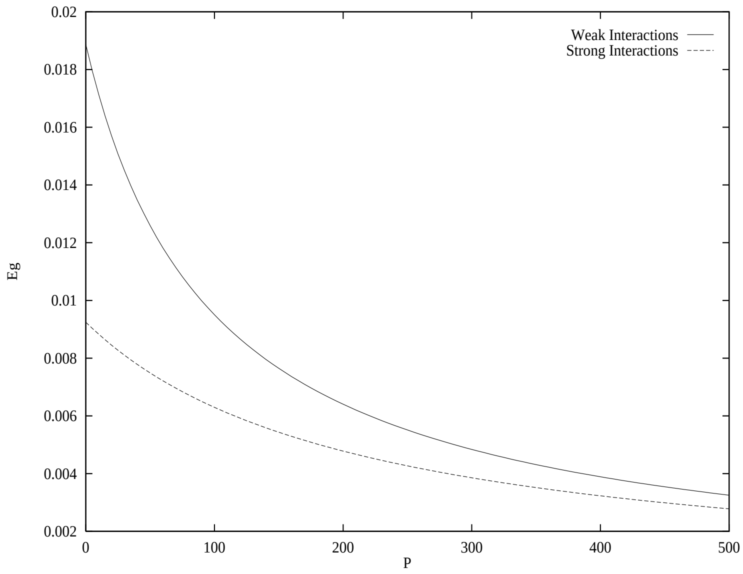


Figure 8: E_G versus number of training pairs for weakly-interacting hidden units (top curve) and strongly-interacting hidden units (bottom curve).

weights, it has been possible analytically to derive expressions for the generalisation error induced by utilising two separate generalisation mechanisms: the Gibbs and Bayesian generalisers. These expressions are generic in that they are independent of the particular dataset employed; instead they indicate the typical performance that can be expected from the RBF architecture.

In the $\gamma \to 0$ limit, in which the distribution of student weight vectors is effectively induced solely by the training data, both measures of generalisation error, E_G and E_B, were found to be inversely proportional to the number of training pairs, P.

The optimal settings of the training parameters γ and β have been examined; it was determined, empirically for the general case and analytically for the simplified situation of no interactions between basis functions, that minimisation of E_B occurs when γ and β are merely set in the correct ratio. However, this result does not apply to E_G, for which each parameter must be optimised separately.

Finally, the interactions between basis functions were shown to be important for rapid learning: strong interactions allow each hidden node to adapt to every training point, while weak interactions imply some training data is effectively ignored by some hidden units.

Much work remains to be performed in understanding learning and generalisation, both specifically in Radial Basis Function networks and in the general case. An investigation of the analytical tractability of reducing the limitations imposed by the assumption of form for the teacher network is currently in progress.

Acknowledgements

The authors wish to thank the anonymous referees for their helpful comments.

Jason Freeman acknowledges the financial support of the Engineering and Physical Sciences Research Council of Great Britain.

Bibliography

Barron, A.R. (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115-133.

Barron, A.R. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory, 39(3):930-945*

Botros, S.M. and Atkeson, C.G. (1991). Generalization Properties of Radial Basis Functions. In *Advances in Neural Information Processing Systems 3*, Lippman, R.P., Moody, J.E. and Touretzky, eds., pp. 707-713. Morgan Kauffman, San Mateo, CA.

Bruce, A.D. and Saad, D, (1994). Statistical mechanics of hypothesis evaluation. *J. Phys. A.* 27(10):3355-3363.

Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica*, 35D:335–356.

Hansen, L.K. (1993). Stochastic linear learning: Exact test and training error averages. *Neural Networks*, 6:393–396.

Hartman, E.J., Keeler, J.D. and Kowalski, J.M. (1990). Layered neural networks with gaussian hidden units as universal approximators. *Neural Computation*, 2:210–215.

Haussler, D. (1994) The probably approximately correct (PAC) and other learning models. In Meyrowitz, A. and Chipman, S. (eds.) *Foundations of Knowledge Acquisition: Machine Learning* Kluwer

Hertz, J., Krogh, A. and Palmer, R.G. (1989). *Introduction to the Theory of Neural Computation*, volume I of *Santa Fe Institute Lecture Notes.* Addison Wesley.

Levin, E., Tishby, N. and Solla, S.A. (1989). A statistical approach to learning and generalisation in layered neural networks. In *Colt '89: 2nd Workshop on Computational Learning Theory*, pages 245–260.

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

Musavi, M.T., Ahmed, W *et al.* . (1992). On the training of radial basis function classifiers. *Neural Networks*, 5(4):595–603.

Niranjan, M. and Fallside, F. (1990). Neural networks and radial basis functions in classifying static speech patterns. *Computer Speech and Language*, 4:275–289.

Niyogi, P and Girosi, F. (1994) On the relationship between generalization error, hypothesis complexity and sample complexity for radial basis functions *Memo No. 1467, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.*

Radons, G., Schuster, H.G. and Werner, D. (1990). Drift and diffusion in back-propagation learning. In Eckmiller, R. *et al*, (eds.), *Parallel Processing in Neural Systems and Computers.* Elsevier Science Publishers, North Holland.

Rögnvaldsson, T. (1994) On Langevin updating in multilayer perceptrons. *Neural Comp.*, 6:916-926

Schwarze, H. (1993). Learning a rule in a multilayer neural network. *J. Phys. A: Math. Gen.*, 26:5781–5794.

Vapnik, V.N. and Chervonenkis, A.Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities *Theory of Probability and its Applications*, 17(2):264-280.

Watkin, T.L.H, Rau, A, Biehl, M. The statistical mechanics of learning a rule. *Reviews of Modern Physics* 65:499-556.