Communicated by John Platt

# Online Learning in Radial Basis Function Networks

**Jason A. S. Freeman**
*Centre for Cognitive Science, University of Edinburgh, Edinburgh EH8 9LW, U.K.*

**David Saad**
*Department of Computer Science and Applied Mathematics, University of Aston, Birmingham B4 7ET, U.K.*

**An analytic investigation of the average case learning and generalization properties of radial basis function (RBFs) networks is presented, utilizing online gradient descent as the learning rule. The analytic method employed allows both the calculation of generalization error and the examination of the internal dynamics of the network. The generalization error and internal dynamics are then used to examine the role of the learning rate and the specialization of the hidden units, which gives insight into decreasing the time required for training. The realizable and some overrealizable cases are studied in detail: the phase of learning in which the hidden units are unspecialized (symmetric phase) and the phase in which asymptotic convergence occurs are analyzed, and their typical properties found. Finally, simulations are performed that strongly confirm the analytic results.**

## 1 Introduction

Several tools facilitate the analytic investigation of learning and generalization in supervised neural networks, such as the statistical physics methods (see Watkin, Rau, & Biehl, 1993, for a review), the Bayesian framework (MacKay, 1992), and the "probably approximately correct" (PAC) method (Haussler, 1994). These tools have principally been applied to simple networks, such as linear and boolean perceptrons, and various simplifications of the committee machine (see, for instance, Schwarze, 1993, and references therein). It has proved very difficult to obtain general results for the commonly used multilayer networks, such as the sigmoid multilayer perceptron (MLP) and the radial basis function (RBF) network.

Another approach, based on studying the dynamics of online gradient descent training scenarios, has been used by several authors (Heskes & Kappen, 1991; Leen & Orr, 1994; Amari, 1993) to examine the evolution of system parameters, primarily in the asymptotic regime. A similar approach, based on examining the dynamics of overlaps between characteristic sys-

 

tem vectors in online training scenarios, has been suggested recently (Saad & Solla, 1995a, 1995b) for investigating the learning dynamics in the soft committee machine (SCM) (Biehl & Schwarze, 1995). This approach provides a complete description of the learning process, formulated in terms of the overlaps between vectors in the system, and it can be easily extended to include general two-layer networks (Riegler & Biehl, 1995).

For RBFs, some analytic studies focus primarily on generalization error. In Freeman and Saad (1995a, 1995b), average case analyses are performed employing a Bayesian framework to study RBFs under a stochastic training paradigm. In Niyogi and Girosi (1994), a bound on generalization error is derived under the assumption that the training algorithm finds a globally optimal solution. Details of studies of RBFs from the perspective of the PAC framework can be found in Holden and Rayner (1995) and its references. These methods focus on a training scenario in which a model is trained on a fixed set of examples using a stochastic training method.

This article presents a method for analyzing the behavior of RBFs in an online learning scenario whereby network parameters are modified after each presentation of an example, which allows the calculation of generalization error as a function of a set of variables characterizing the properties of the adaptive parameters of the network. The dynamical evolution of these variables in the average case can be found, allowing not only the investigation of generalization ability but also the internal dynamics of the network, such as specialization of hidden units, to be analyzed. This tool has previously been applied to MLPs (Saad & Solla, 1995a, 1995b; Rieler & Biehl, 1995).

## 2 Training Paradigms for RBF Networks

RBF networks have been successfully employed over the years in many real-world tasks, providing a useful alternative to MLPs. Furthermore, the RBF is a universal approximator for continuous functions given a sufficient number of hidden units (Hartman, Keeler, & Kowalski, 1990). The RBF architecture consists of a two-layer fully connected network. The mapping performed by each hidden node represents a radially symmetric basis function; within this analysis, the basis functions are considered gaussian, and each is therefore parameterized by two quantities: a vector representing the position of the basis function center in input space and a scalar representing the width of the basis function. For simplicity, the output layer is taken to consist of a single node; this performs a linear combination of the hidden unit outputs.

There are two commonly utilized methods for training RBFs. One approach is to fix the parameters of the hidden layer (both the basis function centers and widths) using an unsupervised technique such as clustering, setting a center on each data point of the training set, or even picking random values (for a review, see Bishop, 1995). Only the hidden-to-output

weights are adaptable, which makes the problem linear in those weights. Although fast to train, this approach results in suboptimal networks since the basis function centers are set to fixed, suboptimal values. The alternative is to adapt the hidden layer parameters—either just the center positions or both center positions and widths. This renders the problem nonlinear in the adaptable parameters, and hence requires an optimization technique, such as gradient descent, to estimate these parameters. The second approach is computationally more expensive but usually leads to greater accuracy of approximation. This article investigates the nonlinear approach in which basis function centers are continuously modified to allow convergence to more optimal models.

There are two methods in use for gradient descent. In *batch learning*, one attempts to minimize the additive training error over the entire data set; adjustments to parameters are performed once the full training set has been presented. The alternative approach, examined here, is *online learning*, in which the adaptive parameters of the network are adjusted after each presentation of a new data point.[1] There has been a resurgence of interest analytically in the online method; technical difficulties caused by the variety of ways in which a training set of given size can be selected are avoided, so complicated techniques such as the replica method (Hertz, Krogh, & Palmer, 1989) are unnecessary.

## 3 Online Learning in RBF Networks

We examine a gradient descent online training scenario on a continuous error measure. The trained model (student) is an RBF network consisting of $K$ basis functions. The center of student basis function (SBF) $b$ is denoted by $\mathbf{m}_b$, and the hidden-to-output weights of the student are represented by $\mathbf{w}$. Training examples will consist of input-output pairs $(\boldsymbol{\xi}, \zeta)$. The components of $\boldsymbol{\xi}$ are uncorrelated gaussian random variables of mean 0, variance $\sigma_\xi^2$, while $\zeta$ is generated by applying $\boldsymbol{\xi}$ to a deterministic teacher RBF, but one in which the number $M$ and the position of the hidden units need not correspond to that of the student, which allows investigation of overrealizable and unrealizable cases.[2] The mapping implemented by the teacher is denoted by $f_T$ and that of the student by $f_S$. The hidden-to-output weights of the teacher are $\mathbf{w}^0$, while the center of teacher basis function $u$ is given by $\mathbf{n}_u$. The vector of SBF responses to input vector $\boldsymbol{\xi}$ is represented by $\mathbf{s}(\boldsymbol{\xi})$, and those of the teacher are denoted by $\mathbf{t}(\boldsymbol{\xi})$. The overall functions computed by

[1] Obviously one may employ a method that is a compromise between the two extremes

[2] This represents a general training scenario since, being universal approximators, RBF networks can approximate any continuous mapping to a desired degree.

the networks are therefore:[3]

$$f_S(\boldsymbol{\xi}) = \sum_{b=1}^{K} w_b \exp\left(-\frac{\|\boldsymbol{\xi} - \mathbf{m}_b\|^2}{2\sigma_B^2}\right) = \mathbf{w} \cdot \mathbf{s}(\boldsymbol{\xi}) \tag{3.1}$$

$$f_T(\boldsymbol{\xi}) = \sum_{u=1}^{M} w_u^0 \exp\left(-\frac{\|\boldsymbol{\xi} - \mathbf{n}_u\|^2}{2\sigma_B^2}\right) = \mathbf{w}^0 \cdot \mathbf{t}(\boldsymbol{\xi}) \tag{3.2}$$

$N$ will denote the dimensionality of input space and $P$ the number of examples presented.

The centers of the basis functions (input-to-hidden weights) and the hidden-to-output weights are considered adjustable; for simplicity, the widths of the basis functions are fixed to a common value $\sigma_B$. The evolution of the centers of the basis functions is described in terms of the overlaps $Q_{bc} \equiv \mathbf{m}_b \cdot \mathbf{m}_c$, $R_{bu} \equiv \mathbf{m}_b \cdot \mathbf{n}_u$, and $T_{uv} \equiv \mathbf{n}_u \cdot \mathbf{n}_v$, where $T_{uv}$ is constant and describes characteristics of the task to be learned.

Previous work in this area (Biehl & Schwarze, 1995; Saad & Solla, 1995a, 1995b; Riegler & Biehl, 1995) has relied on the thermodynamic limit.[4] This limit allows one to ignore fluctuations in the updates of the means of the overlaps due to the randomness of the training examples, and permits the difference equations of gradient descent to be considered as differential equations. The thermodynamic limit is hugely artificial for local RBFs; as the activation is localized, the $N \to \infty$ limit implies that a basis function responds only in the vanishingly unlikely event that an input point falls exactly on its center; there is no obvious reasonable rescaling of the basis functions.[5] The price paid for not taking this limit is that one has no a priori justification for ignoring the fluctuations in the update of the adaptive parameters due to the randomness of the training example. In this work, we calculate both the means and variances of the adaptive parameters, showing that the fluctuations are practically negligible (see section 5).

**3.1 Calculating the Generalization Error.** Generalization error measures the average dissimilarity over input space between the desired mapping $f_T$

[3] Indices $b$, $c$, $d$, and $e$ will always represent SBFs, $u$ and $v$ will represent those of the teacher

[4] $P \to \infty$, $N \to \infty$, and $P/N = \alpha$, where $\alpha$ is finite.

[5] For instance, utilizing

$$\exp\left(-\frac{\|\xi - \mathbf{m}_b\|^2}{2N\sigma_B^2}\right)$$

eliminates all directional information as the cross-term $\xi \cdot \mathbf{m}_b$ vanishes in the thermodynamic limit.

and that implemented by the learning model $f_S$. This dissimilarity is taken as quadratic deviation:

$$E_G = \left\langle \frac{1}{2}\left[f_S - f_T\right]^2 \right\rangle, \tag{3.3}$$

where $\langle \cdots \rangle$ denotes an average over input space with respect to the measure $p(\boldsymbol{\xi})$. Substituting the definitions of equations 3.1 and 3.2 into this leads to:

$$E_G = \frac{1}{2}\left\{ \sum_{bc} w_b w_c \langle s_b s_c \rangle + \sum_{uv} w_u^0 w_v^0 \langle t_u t_v \rangle - 2\sum_{bu} w_b w_u^0 \langle s_b t_u \rangle \right\}. \tag{3.4}$$

Since the input distribution is gaussian, the averages are gaussian integrals and can be performed analytically; the resulting expression for generalization error is given in the appendix. Each average has dependence on combinations of $\boldsymbol{Q}$, $\boldsymbol{R}$, and $\boldsymbol{T}$ depending on whether the averaged basis functions belong to student or teacher.

**3.2 System Dynamics.** Expressions for the time evolution of the overlaps $\boldsymbol{Q}$ and $\boldsymbol{R}$ can be derived by employing the gradient descent rule,

$$\mathbf{m}_b^{p+1} = \mathbf{m}_b^p + \frac{\eta}{N\sigma_B^2}\delta_b(\boldsymbol{\xi} - \mathbf{m}_b),$$

where $\delta_b = (f_T - f_S)w_b s_b$ and $\eta$ is the learning rate, which is explicitly scaled with $1/N$:

$$\langle \Delta Q_{bc} \rangle = \frac{\eta}{N\sigma_B^2}\left\langle \left[\delta_b(\boldsymbol{\xi} - \mathbf{m}_b^p)\cdot \mathbf{m}_c^p + \delta_c(\boldsymbol{\xi} - \mathbf{m}_c^p)\cdot \mathbf{m}_b^p\right]\right\rangle + \left(\frac{\eta}{N\sigma_B^2}\right)^2 \left\langle \delta_b\delta_c(\boldsymbol{\xi} - \mathbf{m}_b^p)\cdot(\boldsymbol{\xi} - \mathbf{m}_c^p)\right\rangle \tag{3.5}$$

$$\langle \Delta R_{bu} \rangle = \frac{\eta}{N\sigma_B^2}\langle \delta_b(\boldsymbol{\xi} - \mathbf{m}_b^p)\cdot \mathbf{n}_u \rangle. \tag{3.6}$$

The hidden-to-output weights can be treated similarly, but here the learning rate is scaled with $1/K$, yielding:[6]

$$\langle \Delta w_b \rangle = \frac{\eta}{K}\langle (f_T - f_S)s_b \rangle. \tag{3.7}$$

These averages are again gaussian integrals, so they can be carried out analytically. The averaged expressions for $\Delta\boldsymbol{Q}$, $\Delta\boldsymbol{R}$, and $\Delta w$ are given in the appendix.

[6] For simplicity we use the same learning rate for both the centers and the hidden-to-output weights, although different learning rates may be employed.

By iterating equations 3.5, 3.6, and 3.7, the evolution of the learning process can be tracked. This allows one to examine facets of learning such as specialization of the hidden units. Since generalization error depends on $Q$, $R$, and $w$, one can also use these equations with equation 3.4 to track the evolution of generalization error.

## 4 Analysis of Learning Scenarios

**4.1 The Evolution of the Learning Process.** Solving the difference equations 3.5, 3.6, and 3.7 iteratively, one obtains solutions to the mean behavior of the overlaps and the weights. There are four distinct phases in the learning process, which are described with reference to an example of learning an exactly realizable task. The task consists of three SBFs learning a graded teacher of three teacher basis functions (TBFs) where graded implies that the square norms of the TBFs (diagonals of $T$) differ from one another; for this task, $T_{00} = 0.5$, $T_{11} = 1.0$, and $T_{22} = 1.5$.

In this demonstration the teacher is chosen to be uncorrelated, so the off-diagonals of $T$ are 0, and the teacher hidden-to-output weights $w^0$ are set to 1. The learning process is illustrated in Figure 1. Figure 1a (solid curve) shows the evolution of generalization error, calculated from equation 3.4, while Figures 1b–d show the evolution of the equations for the means of $R$, $Q$, and $w$, respectively, calculated by iterating equations 3.5, 3.6, and 3.7 from random initial conditions sampled from the following uniform distributions: $Q_{bb}$ and $w_b$ are sampled from $U[0, 0.1]$, while $Q_{bc,b\neq c}$ and $R_{bc}$ from a uniform distribution $U[0, 10^{-6}]$. These initial conditions will be used throughout the article and reflect random correlations expected by arbitrary initialization of large systems. Input dimensionality $N = 8$, learning rate $\eta = 0.9$, input variance $\sigma_\xi^2 = 1$, and basis function width $\sigma_B^2 = 1$ will be employed unless stated otherwise.

Initially, there is a short transient phase in which the overlaps and hidden-to-output weights evolve from their initial conditions until they reach an approximately steady value ($P = 0$ to $P = 1000$). The symmetric phase then begins, which is characterized by a plateau in the evolution of the generalization error (see Figure 1a, solid curve; $P = 1000$ to $P = 7000$), corresponding to a lack of differentiation among the hidden units; they are unspecialized and learn an average of the hidden units of the teacher, so that the student center vectors and hidden-to-output weights are similar (see Figures 1b–d). The difference in value between the overlaps $R$ between student center vectors and teacher center vectors (see Figure 1b) is only due to the difference in the lengths of various teacher center vectors; if the overlaps were normalized, they would be identical. The symmetric phase is followed by a symmetry-breaking phase in which the SBFs learn to specialize, and become differentiated from one another ($P = 7000$ to $P = 20{,}000$). Finally there is a long convergence phase, as the overlaps and hidden-to-output weights reach their asymptotic values. Since the task is realizable,

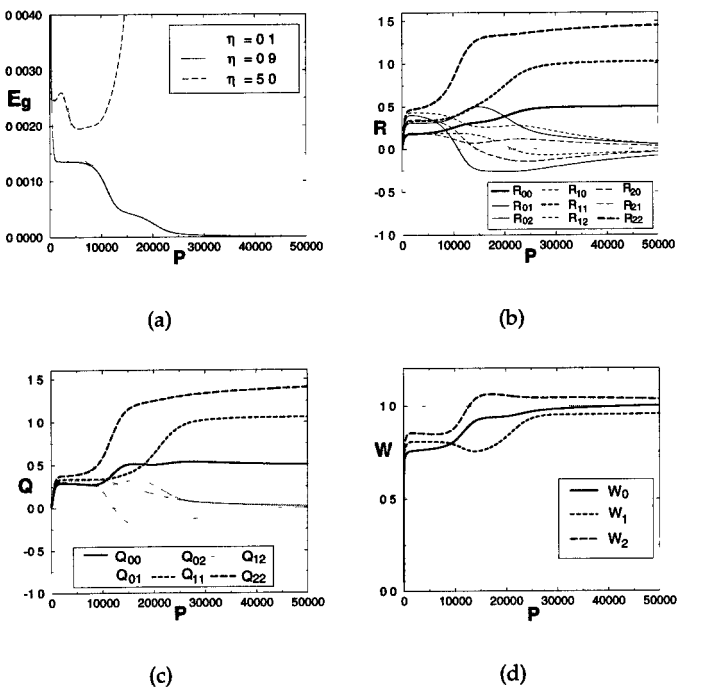


Figure 1: The exactly realizable scenario with positive TBFs. Three SBFs learn a graded, uncorrelated teacher of three TBFs with $T_{00} = 0.5$, $T_{11} = 1.0$, and $T_{22} = 1.5$. All teacher hidden-to-output weights are set to 1. (a) The evolution of the generalization error as a function of the number of examples for several different learning rates ($\eta = 0.1, 0.9, 5.0$). (b, c) The evolution of overlaps between student and teacher center vectors and among student center vectors, respectively. (d) The evolution of the mean hidden-to-output weights.

this phase is characterized by $E_g \to 0$ (see Figure 1a, solid curve) and by the student center vectors and hidden-to-output weights approaching those of the teacher (i.e., $Q_{00} = R_{00} = 0.5, Q_{11} = R_{11} = 1.0, Q_{22} = R_{22} = 1.5$, with the off-diagonal elements of both $Q$ and $R$ being zero; $\forall b, w_b = 1$).[7]

These phases are generic in that they are observed, sometimes with some variation such as a series of symmetric and symmetry-breaking phases, in every online learning scenario for RBFs so far examined. They also correspond to the phases found for MLPs (Saad & Solla, 1995b; Riegler & Biehl, 1995).

[7] The arbitrary labels of the SBFs were permuted to match those of the teacher.

The formalism describes the evolution of the means (and the variances) from certain initial conditions. Convergence of the dynamics to suboptimal attractive fixed points (local minima) may occur if the starting point is within the corresponding basin of attraction. No local minima have been observed in our solutions, which may be an artifact of the system dimensionality.

**4.2 The Role of the Learning Rate.** With all the TBFs positive, analysis of the time evolution of the generalization error, overlaps, and hidden-to-output weights for various settings of the learning rate reveal the existence of three distinct behaviors. If $\eta$ is chosen to be too small (here, $\eta = 0.1$), there is a long period in which there is no specialization of the SBFs and no improvement in generalization ability. The process becomes trapped in a symmetric subspace of solutions; this is the symmetric phase. Given asymmetry in the student initial conditions (in $\boldsymbol{R}$, $\boldsymbol{Q}$, or $\mathbf{w}$) or of the task itself, this subspace will always be escaped, but the time period required may be prohibitively large (see Figure 1a, dotted curve). The length of the symmetric phase increases with the symmetry of the initial conditions. At the other extreme, if $\eta$ is set too large, an initial transient takes place quickly, but there comes a point from which the student vector norms grow extremely rapidly, until the point where, due to the finite variance of the input distribution and local nature of the basis functions, the SBFs are no longer activated during training (see Figure 1a, dashed curve, with $\eta = 5.0$). In this case, the generalization error approaches a finite value as $P \to \infty$, and the task is not solved. Between these extremes lies a region in which the symmetric subspace is escaped quickly, and $E_G \to 0$ as $P \to \infty$ for the realizable case (see Figure 1a, solid curve, with $\eta = 0.9$). The SBFs become specialized and, asymptotically, the teacher is emulated exactly.

These results for the learning rate are qualitatively similar to those found for SCMs and MLPs (Biehl & Schwarze, 1995; Saad & Solla, 1995a, 1995b; Riegler & Biehl, 1995).

**4.3 Task Dependence.** The symmetric phase depends on the symmetry of the task as well as that of the initial conditions. One would expect a shorter symmetric phase in inherently asymmetric tasks. To examine this, a task similar to that of section 4.1 was employed, with the single change being that the sign of one of the teacher hidden-to-output weights was flipped, thus providing two categories of targets: positive and negative. The initial conditions of the student remained the same as in the previous task, with $\eta = 0.9$.

The evolution of generalization error and the overlaps for this task are shown in Figure 2. The dividing of the targets into two categories effectively eliminates the symmetric phase; this can be seen by comparing the evolution of the generalization error for this task (see Figure 2a, dashed curve) with that for the previous task (see Figure 2a, solid curve). There is no longer a plateau in the generalization error. Correspondingly, the symmetries be-

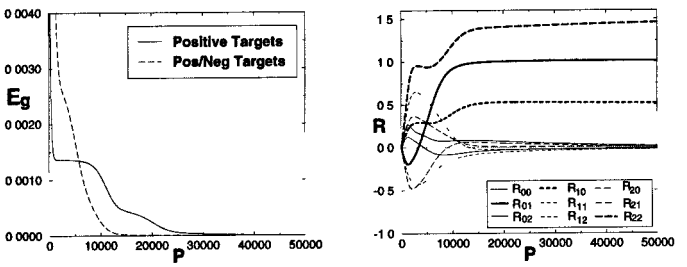


Figure 2: The exactly realizable scenario defined by a teacher network with a mixture of positive and negative TBFs. Three SBFs learn a graded, uncorrelated teacher of three TBFs with $T_{00} = 0.5, T_{11} = 1.0$, and $T_{22} = 1.5$. $w_0^0 = 1, w_1^0 = -1, w_2^0 = 1$. (a) The evolution of the generalization error for this case and, for comparison, the evolution in the case of all positive TBFs. (b) The evolution of the overlaps between student and teacher centers $R$.

tween SBFs break immediately, as can be seen by examining the overlaps between student and teacher center vectors (see Figure 2b); this should be compared with figure 1b, which denotes the evolution of the overlaps in the previous task. Note that the plateaus in the overlaps (see Figure 1b, $P = 1000$ to $P = 7000$) are not found for the antisymmetric task.

The elimination of the symmetric phase is an extreme result caused by the extremely asymmetric teacher. For networks with many hidden units, one can find a cascade of subsymmetric phases, each shorter than the single symmetric phase in the corresponding task with only positive targets, in which there is one symmetry between the hidden units seeking positive targets and another between those seeking negative targets.

This suggests a simple and easily implemented strategy for increasing the speed of learning when targets are predominantly positive (negative): eliminate the bias of the training set by subtracting (adding) the mean target from each target point. This corresponds to an old heuristic among RBF practitioners. It follows that the hidden-to-output weights should be initialized from a zero-mean distribution.

**4.4 The Overrealizable Case.** In real-world problems, the exact form of the data-generating mechanism is rarely known. This leads to the possibility that the student may be overly powerful, in that it is capable of fitting surfaces more complicated than that of the true teacher. It is important to gain insight into how architectures will respond given such a scenario

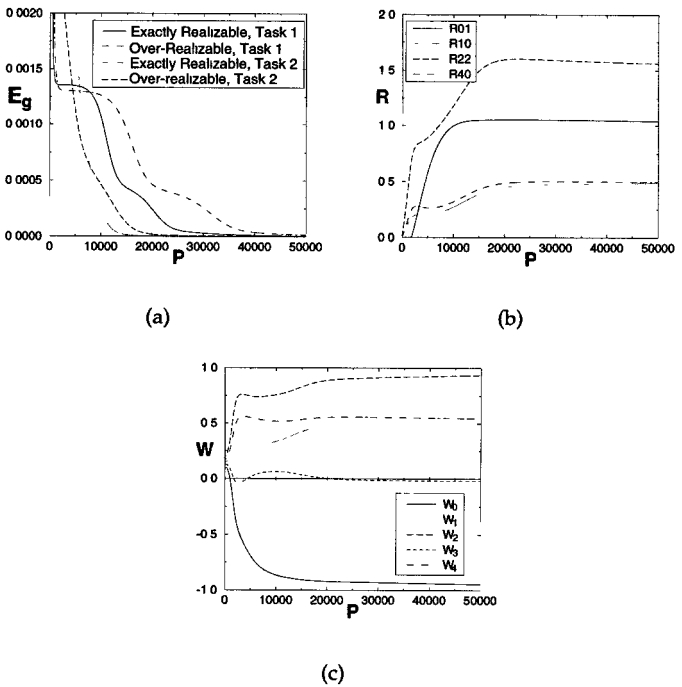


Figure 3: The overrealizable scenario. (a) The evolution of the generalization error in two tasks; each task is learned by a well-matched student (exactly realizable) and an overly powerful student (overrealizable). (b, c) The evolution of the overlaps $R$ and the hidden-to-output weights $\mathbf{w}$ for the overrealizable case in the second task, in which the teacher RBF includes a mixture of positive and negative hidden-to-output weights. In this scenario, five SBFs learn a graded, uncorrelated teacher of three TBFs with $T_{00} = 0.5, T_{11} = 1.0$, and $T_{22} = 1.5$. $w_0^0 = 1, w_1^0 = -1, w_2^0 = 1$.

in order to be confident that they can be used successfully when the true teacher is unknown.

Intuitively, one might expect that a student that is well matched to the teacher will learn faster than one that is overly powerful. Figure 3a shows two tasks, each of which compares the overrealizable scenario with the well-matched case. The first task, consisting of three TBFs, is identical to that detailed in section 4.1, and hence has only positive targets. The performance of a well-matched student of three SBFs is compared with an overrealizable scenario in which five SBFs learn the three TBFs. Comparison of the evolution of generalization error between these learning scenarios

is shown in Figure 3a; the solid curve represents the well-matched scenario, and the dot-dash curve illustrates the overrealizable scenario. The length of the symmetric phase is significantly increased with the overly powerful student. The length of the convergence phase is also increased. An analytical treatment of these effects as well as the overrealizable scenario generally is given elsewhere (Freeman & Saad, in press).

The second task deals with the alternative scenario in which one TBF has a negative hidden-to-output weight; the task is identical to that defined in section 4.3, and the student initial conditions are again as specified in section 4.1. In Figure 3a the evolution of generalization error for both the overrealizable scenario (dashed curve) in which five SBFs learn three TBFs, and the corresponding well-matched case in which three SBFs learn three TBFs (dotted curve), is shown. There is no well-defined symmetric case due to the inherent asymmetry of the task. The convergence phase is again greatly increased in length; this appears to be a general feature of the overrealizable scenario.

Given that the student is overly powerful, there appear to be, a priori, several remedies available to the student: eliminate the excess nodes, form cancellation pairs (in which two students exactly cancel one another), or devise more complicated fitting schemes.

To examine the actual responses of the student, the evolution of the overlaps between student and teacher and of the hidden-to-output weights for the particular scenario described by the second trial detailed is presented in Figures 3b and 3c, respectively. Looking first at Figure 3c, it is apparent that $w_3$ approaches zero (short-dashed curve), indicating that SBF 3 is entirely eliminated during training. Thus four SBFs remain to emulate three TBFs. The negative TBF 1 is exactly emulated by SBF 0, as $T_{11} = 1, w_1^0 = -1$, and $R_{01} = 1, w_0 = -1$ (solid curve on both Figures 3b and 3c), while, similarly, SBF 2 exactly emulates TBF 2 (long-dashed curve, both figures). This leaves SBF 1 and SBF 4 to emulate TBF 0. Looking at Figure 3c, dotted and dot-dash curves, both student hidden-to-output weights approach 0.5, exactly half that of the hidden-to-output weight of TBF 0; looking at Figure 3b, both SBFs have 0.5 overlap with TBF 0. This indicates that the sum of both students emulates TBF 0. Thus, elimination and fitting involving the noncancelling combination of nodes were found; in these trials and many others, no pairwise cancellation was found. One presumes that this could be induced by very careful selection of the initial conditions but that it is not found under normal circumstances.

**4.5 Analysis of the Symmetric Phase.** The symmetric phase, in which there is no specialization of the hidden units, can be analyzed in the realizable case by employing a few simplifying assumptions. It is a phenomenon that is predominantly associated with small $\eta$, so terms of $\eta^2$ are neglected. The hidden-to-output weights are clamped to +1. The teacher is taken to be isotropic: TBF centers have identical norms of 1, each hav-

ing no overlap with the others; therefore $T_{uv} = \delta_{uv}$. This has the result that the student norms $Q_{bb}$ are very similar in this phase, as are the student-student correlations, so $Q_{bb} \equiv Q$ and $Q_{bc,b\neq c} \equiv C$, where $Q$ becomes the square norm of the SBFs, and $C$ is the overlap between any two different SBFs.

Following the geometric argument of Saad & Solla (1995b), in the symmetric phase, the SBF centers are confined to the subspace spanned by the TBF centers. Since $T_{uv} = \delta_{uv}$, the SBF centers can be written in the orthonormal basis defined by the TBF centers, with the components being the overlaps $\boldsymbol{R}$: $\boldsymbol{m}_b = \sum_{u=1}^{M} R_{bu}\boldsymbol{n}_u$. Because the teacher is isotropic, the overlaps are independent of both $b$ and $u$ and thus can be written in terms of a single parameter $R$. Further, this reduction to a single overlap parameter leads to $Q = C = MR^2$, so the evolution of the overlaps can be described as a single difference equation for $R$. The analytic solution of equations 3.5, 3.6, and 3.7 under these restrictions is still rather complicated. However, since we are primarily interested in large systems, that is, large $K$, we examine the dominant terms in the solution. Expanding in $1/K$ and discarding second-order terms renders the system simple enough to solve analytically for the symmetric fixed point:

$$R = \frac{1}{K\left(1 + \sigma_B^2 - \sigma_B^2 \exp\left[\left(\frac{1}{2\sigma_B^2}\right)\frac{\sigma_B^2+1}{\sigma_B^2+2}\right]\right)}. \tag{4.1}$$

The stability of the fixed point, and thus the breaking of the symmetric phase, can be examined by an eigenvalue analysis of the dynamics of the system near the fixed point. The method employed is similar to that detailed in Saad and Solla (1995b) and is presented elsewhere (Freeman & Saad, in press). The dominant eigenvalue ($\lambda_1 > 0$) scales with $K$ and represents a perturbation that breaks the symmetries between the hidden units; the remaining modes $\lambda_{i\neq 1} < 0$, which also scale with $K$, are irrelevant because they preserve the symmetry. This result is in contrast to that for the SCM (Saad & Solla, 1995b), in which the dominant eigenvalue scales with $1/K$. This implies that for RBFs, the more hidden units in the network, the faster the symmetric phase is escaped, resulting in negligible symmetric phases for large systems, while in SCMs the opposite is true. This difference is caused by the contrast between the localized nature of the basis function in the RBF network and the global nature of sigmoidal hidden nodes in SCM. In the SCM case, small perturbations around the symmetric fixed point result in relatively small changes in error since the sigmoidal response changes very slowly as one modifies the weight vectors. On the other hand, the gaussian response decays exponentially as one moves away from the center, so small perturbations around the symmetric fixed point result in massive changes that drive the symmetry breaking. When $K$ increases, the

error surface looks very rugged, emphasizing the peaks and increasing this effect, in contrast to the SCM case, where more sigmoids means a smoother error surface.

This does not mean that the symmetric phase can be ignored for realistically sized networks, however. Even with a teacher that is not particularly symmetric, this phase can play a significant role in the learning dynamics. To demonstrate this, a teacher RBF of 10 hidden units with $N = 5$ was constructed with the teacher centers generated from a gaussian distribution $\mathcal{N}[0, 0.5]$. Note that this teacher must be correlated because the number of centers is larger than the input dimension. A student network, also of 10 hidden units, was constructed with all weights initialized from $\mathcal{N}[0, 0.05]$. The networks were then mapped into the corresponding overlaps, and the learning process was run with $\eta = 0.1$. The evolution of generalization error is shown in Figure 4d: the symmetric phase, extending here from $P = 2000$ to $P = 15{,}000$, is a prominent phenomenon of the learning dynamics. It is not merely an artifact of a highly symmetric teacher configuration (the teacher was random and correlated) or of a specially chosen set of initial conditions, as the student was initialized with realistic initial conditions before being mapped into overlaps.

**4.6 Analysis of the Convergence Phase.** To gain insight into the convergence of the online gradient descent process in a realizable scenario, a similar simplified learning scenario to that used in the symmetric phase analysis was employed. The hidden-to-output weights are again fixed to $+1$, and the teacher is defined by $T_{uv} = \delta_{uv}$. The scenario can be extended to adaptable hidden-to-output weights (this is presented in Freeman & Saad, in press, along with more mathematical detail). As in the symmetric phase, the fact that $T_{uv} = \delta_{uv}$ allows the system to be reduced to four adaptive quantities: $Q \equiv Q_{bb}$, $C \equiv Q_{bc,b\neq c}$, $R \equiv R_{bb}$, and $S \equiv R_{bc,b\neq c}$.

Linearizing this system about the known fixed point of the dynamics, $Q = 1, C = 0, R = 1, S = 0$, yields an equation of the form $\Delta \mathbf{x} = A\mathbf{x}$, where $\mathbf{x} = \{1 - R, 1 - Q, S, C\}$ is the vector of deviations from the fixed point. The eigenvalues of the matrix $A$ control the converging system; these are presented in Figure 4a for $K = 10$. In every case examined, there is a single critical eigenvalue $\lambda_c$ that controls the stability and convergence rate of the system (shown in bold), a nonlinear subcritical eigenvalue, and two subcritical linear eigenvalues. The value of $\eta$ at $\lambda_c = 0$ determines the maximum learning rate for convergence to occur; for $\lambda_c > 0$ the fixed point is unstable. The convergence of the overlaps is controlled by the critical eigenvalue; therefore, the value of $\eta$ at the single minimum of $\lambda_c$ determines the optimal learning rate ($\eta_{\text{opt}}$) in terms of the fastest convergence of the system to the fixed point.

Examining $\eta_c$ and $\eta_{\text{opt}}$ as a function of $K$ (see Figure 4b), one finds that both quantities scale as $1/K$; the maximum and optimal learning rates are

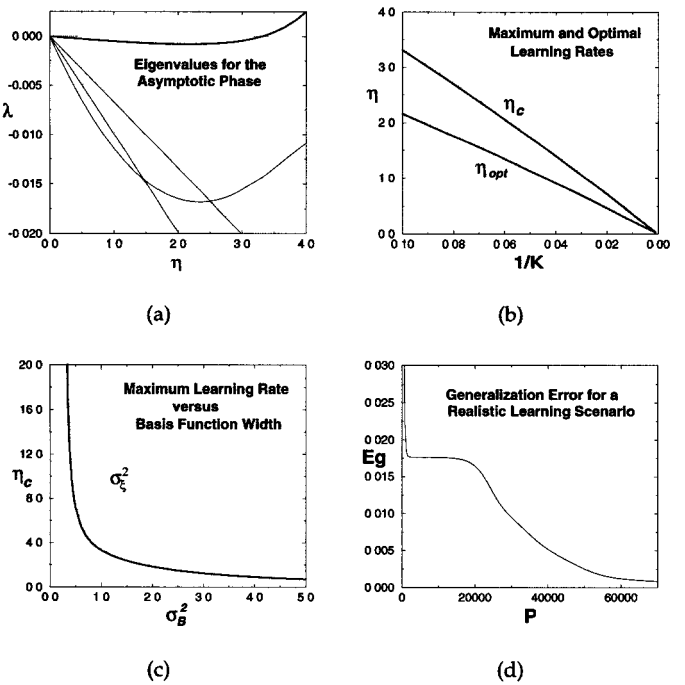


Figure 4: Convergence and symmetric phases. (a) The eigenvalues controlling the dynamics of the system for the convergence phase (detailed in section 4.6), linearized about the asymptotic fixed point in the realizable case, as a function of $\eta$. The critical eigenvalue is shown in bold. (b) The maximum and optimal convergence phase learning rates, found from the critical eigenvalue; these quantities scale as $1/K$. (c) The maximum convergence phase learning rate as a function of basis function width. (d) The evolution of generalization error for a realistically sized learning scenario (described in section 4.5), demonstrating that the symmetric phase can play a significant role, even with a correlated, asymmetric teacher.

inversely proportional to the number of hidden units of the student. Numerically, the ratio of $\eta_{\text{opt}}$ to $\eta_c$ is approximately two-thirds.

Finally, the relationship between basis function width and $\eta_c$ is plotted in Figure 4c. When the widths are small, $\eta_c$ is very large as it becomes unlikely that a training point will activate any of the basis functions. For $\sigma_B^2 > \sigma_\xi^2$, $\eta_c \sim 1/\sigma_B^2$.

## 5 Quantifying the Variances

Because the thermodynamic limit is not employed, it is necessary to quantify the variances in the adaptive parameters to justify considering only the mean updates.[8]

By making assumptions as to the form of these variances, it is possible to derive equations describing their evolution. Specifically, it is assumed that each update function and parameter being updated can be written in terms of a mean and fluctuation; for instance, applying this to $Q_{bc}$:

$$\Delta Q_{bc} = \overline{\Delta Q_{bc}} + \widehat{\Delta Q_{bc}} \quad Q_{bc} = \overline{Q_{bc}} + \sqrt{\frac{\eta}{N}}\widehat{Q_{bc}}, \tag{5.1}$$

where $\langle Q_{bc} \rangle$ denotes an average value and $\widehat{Q_{bc}}$ represents a fluctuation due to the randomness of the example. Combining these equations and averaging with respect to the input distribution results in a set of difference equations describing the evolution of the variances of the overlaps and hidden-to-output weights (similar to Riegler & Biehl, 1995) as training proceeds. Details of the method can be found in Heskes and Kappen (1991) and in Barber, Saad, and Sollich (1996) for the SCM. It has been shown that the variances vanish in the thermodynamic limit for realizable cases (Barber et al., 1996; Heskes & Kappen, 1991). (A detailed description of the calculation of the variances as applied to RBFs appears in Freeman & Saad, in press.)

Figure 5 shows the evolution of the variances, as error bars on the mean, for the dominant overlaps and the hidden-to-output weights using $\eta = 0.9, N = 10$ on a task identical to that described in section 4.1. Examining the dominant overlaps $R$ first (see Figure 5a), the variances follow the same pattern for each overlap but at different values of $P$. The variances begin at 0, then increase, peaking at the symmetry-breaking point at which the SBF begins to specialize on a particular TBF; then they decrease to 0 again as convergence occurs. Looking at each SBF in turn, for SBF 2 (dashed curve), the overlap begins to specialize at approximately $P = 2000$, where the variance peak occurs; for SBF 0 (solid curve), the symmetry lasts until $P = 10{,}000$, again where the variance peak occurs, and for SBF 1 (dotted curve), the symmetry breaks later at approximately $P = 20{,}000$, again where the peak of the variance occurs. The variances then dwindle to 0 for each SBF in the convergence phase.

Essentially the same pattern occurs for the hidden-to-output weights (see Figure 5b). The variances increase rapidly until the hidden units begin

[8] The hidden-to-output weights are not intrinsically self-averaging even in the thermodynamic limit, although they have been shown to be such for the MLP if the learning rate is scaled with $N$ (Riegler & Biehl, 1995). If scaled differently, adiabatic elimination techniques may be employed to describe the evolution adequately (Riegler, personal communication, 1996).

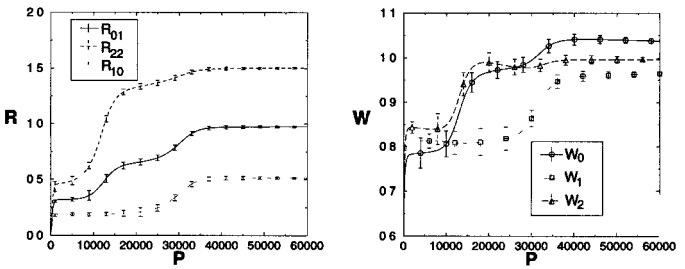


Figure 5: Evolution of the variances of the overlaps $R$ and hidden-to-output weights $w$ are shown in (a) and (b), respectively. The curves denote the evolution of the means; the error bars show the evolution of the fluctuations about the mean. Input dimensionality $N = 10$, learning rate $\eta = 0.9$, input variance $\sigma_\xi^2 = 1$, and basis function width $\sigma_B^2 = 1.0$.

to specialize, at which point the variances peak. This is followed by the variances' decreasing to 0 as convergence occurs. For both overlaps and hidden-to-output weights, the mean is an order of magnitude larger than the standard deviation at the variance peak and is much more dominant elsewhere; the ratio becomes greater as $N$ is increased.

The magnitude of the variances is influenced by the degree of symmetry of the initial conditions of the student and the task in that the greater this symmetry is, the larger the variances. Discussion of this phenomenon can be found in Barber et al. (1996); it will be explored at greater length for RBFs in a future publication.

## 6 Simulations

In order to confirm the validity of the analytic results, simulations were performed in which RBFs were trained using online gradient descent. The trajectories of the overlaps were calculated from the trajectories of the weight vectors of the network, and generalization error was estimated by finding the average error on a 1000-point test set. The procedure was performed 50 times and the results averaged, subject to permutation of the labels of the SBFs to ensure the average was meaningful.

Typical results are shown in Figure 6. The example shown is for an exactly realizable system of three SBFs and three TBFs at $N = 5, \eta = 0.9$. Figure 6a shows the correspondence between empirical test error and theoretical generalization error. At all times, the theoretical result is within

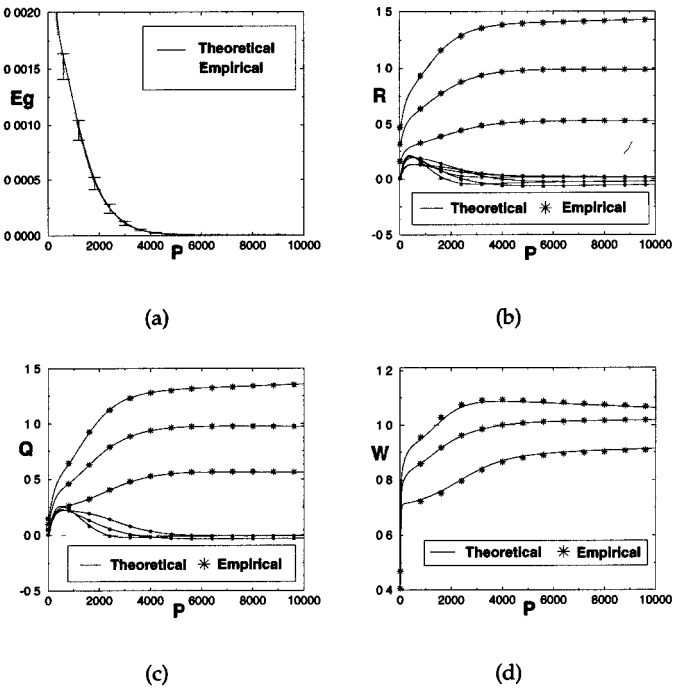


Figure 6: Comparison of theoretical results with simulations. The simulation results are averaged over 50 trials; the labels of the student hidden units were permuted where necessary to make the averages meaningful. Empirical generalization error was approximated with the test error on a 1000-point test set (a). Error bars on the simulations are at most the size of the larger asterisks for the overlaps (b, c), and at most twice this size for the hidden-to-output weights (d). Input dimensionality $N = 5$, learning rate $\eta = 0.9$, input variance $\sigma_\xi^2 = 1$, and basis function width $\sigma_B^2 = 1$.

one standard deviation of the empirical result. Figures 6b–d show the excellent correspondence between the trajectories of the theoretical overlaps and hidden-to-output weights and their empirical counterparts; the error bars on the simulation distributions are not shown because they are approximately the size of the symbols. The simulations demonstrate the validity of the theoretical results. In addition, we have found excellent correlation between the analytically calculated variances and those obtained from the simulations (this is explored further in Freeman & Saad, in press).

## 7 Conclusion

Online learning, in which the adaptive parameters of the network are updated at each presentation of a data point, was examined for the RBF using gradient descent learning. The analytic method presented allows the calculation of the evolution of generalization error and the specialization of the hidden units.

This method was used to elucidate the stages of training and the role of the learning rate. There are four stages of training: a short transitory phase in which the adaptive parameters move from the initial conditions to the symmetric phase; the symmetric phase itself, characterized by lack of differentiation among hidden units; a symmetry-breaking phase in which the hidden units become specialized; and a convergence phase in which the adaptive parameters reach their final values asymptotically. Three regimes were found for the learning rate: small, giving unnecessarily slow learning; intermediate, leading to fast escape from the symmetric phase and convergence to the correct target; and too large, which results in a divergence of SBF norms and failure to converge to the correct target.

Examining the exactly realizable scenario, it was shown that employing both positive and negative targets leads to much faster symmetry breaking; this appears to be the underlying reason behind the neural network folklore that targets should be given zero mean. The overrealizable case was also studied, showing that overrealizability extends both the length of the symmetric phase and that of the convergence phase.

The symmetric phase for realizable scenarios was analyzed and the value of the overlaps at the symmetric fixed point found. It was discovered that there is a significant difference between the behaviors of the RBF and SCM, in that increasing $K$ speeds up the symmetry-breaking in RBFs, while it slows the process for SCMs.

The convergence phase was also studied; both maximum and optimal learning rates were calculated and shown to scale as $1/K$. The dependence of the maximum learning rate on the width of the basis functions was also examined, and, for $\sigma_B^2 > \sigma_\xi^2$, the maximum learning rate scales approximately as $1/\sigma_B^2$.

Finally, simulations were performed that strongly confirm the theoretical results.

Future work includes the study of unrealizable cases, in which the learning rate must decay over time in order to find a stable solution, the study of the effects of noise and regularizers, the extension of the analysis of the convergence phase to fully adaptable hidden-to-output weights, and the use of the theory to aid real-world learning tasks, by, for instance, deliberately breaking the symmetries between SBFs in order to reduce drastically or even eliminate the symmetric phase.

## Appendix

### Generalization Error

$$E_G = \frac{1}{2}\left\{\sum_{bc} w_b w_c I_2(b,c) + \sum_{uv} w_u^0 w_v^0 I_2(u,v) - 2\sum_{bu} w_b w_u^0 I_2(b,u)\right\} \quad \text{(A.1)}$$

### $\Delta Q$, $\Delta R$, and $\Delta w$

$$\begin{aligned}\langle \Delta Q_{bc}\rangle = \frac{\eta}{N\sigma_B^2}\left\{w_b\left[\bar{J}_2(b;c) - Q_{bc}\bar{I}_2(b)\right] + w_c\left[\bar{J}_2(c;b) - Q_{bc}\bar{I}_2(c)\right]\right\} \\ + \left(\frac{\eta}{N\sigma_B^2}\right)^2 w_b w_c\left\{\bar{K}_4(b,c) + Q_{bc}\bar{I}_4(b,c) \right. \\ \left. - \bar{J}_4(b,c;b) - \bar{J}_4(b,c;c)\right\}\end{aligned} \quad \text{(A.2)}$$

$$\langle \Delta R_{bu}\rangle = \frac{\eta}{N\sigma_B^2} w_b\left\{\bar{J}_2(b;u) - R_{bu}\bar{I}_2(b)\right\} \quad \text{(A.3)}$$

$$\langle \Delta w_b\rangle = \frac{\eta}{K}\bar{I}_2(b) \quad \text{(A.4)}$$

### $\bar{I}$, $\bar{J}$, and $\bar{K}$

$$\bar{I}_2(b) = \sum_u w_u^0 I_2(b,u) - \sum_d w_d I_2(b,d) \quad \text{(A.5)}$$

$$\bar{J}_2(b;c) = \sum_u w_u^0 J_2(b,u;c) - \sum_d w_d J_2(b,d;c) \quad \text{(A.6)}$$

$$\begin{aligned}\bar{I}_4(b,c) = \sum_{de} w_d w_e I_4(b,c,d,e) + \sum_{uv} w_u^0 w_v^0 I_4(b,c,u,v) \\ - 2\sum_{du} w_d w_u^0 I_4(b,c,d,u)\end{aligned} \quad \text{(A.7)}$$

$$\begin{aligned}\bar{J}_4(b,c;f) = \sum_{de} w_d w_e J_4(b,c,d,e;f) + \sum_{uv} w_u^0 w_v^0 J_4(b,c,u,v;f) \\ - 2\sum_{du} w_d w_u^0 J_4(b,c,d,u;f)\end{aligned} \quad \text{(A.8)}$$

$$\begin{aligned}\bar{K}_4(b,c) = \sum_{de} w_d w_e K_4(b,c,d,e) + \sum_{uv} w_u^0 w_v^0 K_4(b,c,u,v) \\ - 2\sum_{du} w_d w_u^0 K_4(b,c,d,u)\end{aligned} \quad \text{(A.9)}$$

*I, J,* **and** *K.* In each case, only the quantity corresponding to averaging over SBFs is presented. Each quantity has similar counterparts in which TBFs are substituted for SBFs. For instance, $I_2(b,c) = \langle s_b s_c \rangle$ is presented, and $I_2(u,v) = \langle t_u t_v \rangle$ and $I_2(b,u) = \langle s_b t_u \rangle$ are omitted.

$$I_2(b,c) = (2l_2\sigma_\xi^2)^{-N/2} \times \exp\left[\frac{-Q_{bb} - Q_{cc} + (Q_{bb} + Q_{cc} + 2Q_{bc})/2\sigma_B^2 l_2}{2\sigma_B^2}\right] \tag{A.10}$$

$$J_2(b,c;d) = \left(\frac{Q_{bd} + Q_{cd}}{2l_2\sigma_B^2}\right) I_2(b,c) \tag{A.11}$$

$$I_4(b,c,d,e) = (2l_4\sigma_\xi^2)^{-N/2} \exp\left[\frac{-Q_{bb} - Q_{cc} - Q_{dd} - Q_{ee}}{2\sigma_B^2}\right] \times \exp\left[\frac{Q_{bb}+Q_{cc}+Q_{dd}+Q_{ee}+2(Q_{bc}+Q_{bd}+Q_{be}+Q_{cd}+Q_{ce}+Q_{de})}{4l_4\sigma_B^4}\right] \tag{A.12}$$

$$J_4(b,c,d,e;f) = \left(\frac{Q_{bf} + Q_{cf} + Q_{df} + Q_{ef}}{2l_4\sigma_B^2}\right) I_4(b,c,d,e) \tag{A.13}$$

$$K_4(b,c,d,e) = \left(\frac{2Nl_4\sigma_B^4 + Q_{bb} + Q_{cc} + Q_{dd} + Q_{ee}}{4l_4\sigma_B^4} + \frac{2(Q_{bc} + Q_{bd} + Q_{be} + Q_{cd} + Q_{ce} + Q_{de})}{4l_4^2\sigma_B^4}\right) I_4(b,c,d,e) \tag{A.14}$$

**Other Quantities**

$$l_2 = \frac{2\sigma_\xi^2 + \sigma_B^2}{2\sigma_B^2\sigma_\xi^2} \tag{A.15}$$

$$l_4 = \frac{4\sigma_\xi^2 + \sigma_B^2}{2\sigma_B^2\sigma_\xi^2} \tag{A.16}$$

**Acknowledgments**

We thank Ansgar West and David Barber for useful discussions, and the anonymous referees for their comments. D.S. thanks the Leverhulme Trust for its support (F/250/K).

## References

Amari, S. (1993). Backpropagation and stochastic gradient descent learning. *Neurocomputing, 5*, 185–196.

Barber, D., Saad, D., & Sollich, P. (1996). Finite size effects in online learning of multilayer neural networks. *Euro. Phys. Lett., 34*, 151–156.

Biehl, M., & Schwarze, H. (1995). Learning by online gradient descent. *J. Phys. A: Math. Gen., 28*, 643.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Freeman, J., & Saad, D. (1995a). Learning and generalisation in radial basis function networks. *Neural Computation, 7*, 1000–1020.

Freeman, J., & Saad, D. (1995b). Radial basis function networks: Generalization in overrealizable and unrealizable scenarios. *Neural Networks, 9*, 1521–1529.

Freeman, J., & Saad, D. (in press). The dynamics of on-line learning in radial basis function networks. *Phys. Rev. A*.

Hartman, E., Keeler, J., & Kowalski, J. (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 A. Meyrowitz & S. Chipman (Eds.), *Foundations of knowledge acquisition: Machine learning* (Chap. 9). Boston: Kluwer.

Hertz, J., Krogh, A., & Palmer, R. (1989). *Introduction to the theory of neural computation*. Reading, MA: Addison-Wesley.

Heskes, T., & Kappen, B. (1991). Learning processes in neural networks. *Phys. Rev. A., 44*, 2718–2726.

Holden, S., & Rayner, P. (1995). Generalization and PAC learning: Some new results for the class of generalized single-layer networks. *IEEE Trans. on Neural Networks, 6*(2), 368–380.

Leen, T., & Orr, G. (1994). Optimal stochastic search and adaptive momentum. In J. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems*, (6: 477–484). San Mateo, CA: Morgan Kaufmann.

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

Niyogi, P., & Girosi, F. (1994). *On the relationship between generalization error, hypothesis complexity and sample complexity for radial basis functions* (Tech. Rep.). Cambridge, MA: AI Laboratory, MIT.

Riegler, P., & Biehl, M. (1995). On-line backpropagation in two-layered neural networks. *J. Phys. A: Math. Gen., 28*, L507–513.

Saad, D., & Solla, S. (1995a). Exact solution for online learning in multilayer neural networks. *Phys. Rev. Lett., 74*, 4337–4340.

Saad, D., & Solla, S. (1995b). On-line learning in soft committee machines. *Phys. Rev. E., 52*, 4225–4243.

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

Watkin, T., Rau, A., & Biehl, M. (1993). The statistical mechanics of learning a rule. *Reviews of Modern Physics, 65*, 499–556.

---

Received March 15, 1996; accepted January 3, 1997.