On the Relationship Between Generalization Error, Hypothesis Complexity, and Sample Complexity for Radial Basis Functions[1]

Partha Niyogi and Federico Girosi

Center for Biological and Computational Learning
and
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts, 02139

Abstract

Feedforward networks are a class of regression techniques that can be used to learn to perform some task from a set of examples. The question of generalization of network performance from a finite training set to unseen data is clearly of crucial importance. In this article we first show that the generalization error can be decomposed in two terms: the approximation error, due to the insufficient representational capacity of a finite sized network, and the estimation error, due to insufficient information about the target function because of the finite number of samples. We then consider the problem of approximating functions belonging to certain Sobolev spaces with Gaussian Radial Basis Functions. Using the above mentioned decomposition we bound the generalization error in terms of the number of basis functions and number of examples. While the bound that we derive is specific for Radial Basis Functions, a number of observations deriving from it apply to any approximation technique. Our result also sheds light on ways to choose an appropriate network architecture for a particular problem and the kinds of problems which can be effectively solved with finite resources, i.e., with finite number of parameters and finite amounts of data.

[1] To be published on *Neural Computation*.

1 Introduction

Many problems in learning theory can be effectively modelled as learning an input output mapping on the basis of limited evidence of what this mapping might be. The mapping usually takes the form of some unknown function between two spaces and the evidence is often a set of labelled, noisy, examples i.e., (x, y) pairs which are consistent with this function. On the basis of this data set, the learner tries to infer the true function. The unknown target function is assumed to belong to some class $\mathcal{F}$ (the *concept class*). Typical examples of concept classes are classes of indicator functions, boolean functions, Sobolev spaces etc. The learner is provided with a finite data set. For our purposes we assume that the data is drawn by sampling independently the input/output space $(X \times Y)$ according to some unknown probability distribution. On the basis of this data, the learner then develops a hypothesis (another function belonging to the *hypothesis class* $H \subset \mathcal{F}$) about the identity of the target function. Hypothesis classes could also be of different kinds. For example, they could be classes of boolean functions, polynomials, Multilayer Perceptrons, Radial Basis Functions and so on.

If, as more and more data becomes available, the learner's hypothesis becomes closer and closer to the target and converges to it in the limit, the target is said to be learnable. The error between the learner's hypothesis and the target function is defined to be the *generalization error* and for the target to be learnable the generalization error should go to zero as the data goes to infinity. While learnability is certainly a very desirable quality, it requires the fulfillment of two important criteria.

First, there is the issue of the representational capacity (or *hypothesis complexity*) of the hypothesis class. This must have sufficient power to represent or closely approximate the concept class. Otherwise for some target function $f \in \mathcal{F}$, the best hypothesis h in H might be far away from it. The error that this best hypothesis makes is formalized later as the *approximation error*.

Second, we do not have infinite data but only some finite random sample set from which we construct a hypothesis. This hypothesis constructed from the finite data might be far from the best possible hypothesis, h, resulting in an additional error. This is formalized later as the *estimation error*. The amount of data needed to ensure a small estimation error is referred to as the *sample complexity* of the problem. The hypothesis complexity, the sample complexity and the generalization error are related. If the class H is very large or in other words has high complexity, then for the same estimation error, the sample complexity increases. If the hypothesis complexity is small, the sample complexity is also small but now for the same estimation error the approximation error is high. This point has been developed in terms of the bias-variance trade-off by Geman, Bienenstock, and Doursat (1992). The bias term corresponds to the approximation error, and the variance corresponds to the estimation error. Other authors have discussed this more generally in the statistics literature (Rissanen, 1989; Vapnik, 1982).

The purpose of this paper is two-fold. First, we formalize the problem of learning from examples so as to highlight the relationship between hypothesis complexity, sample complexity and generalization error. Second, we explore this relationship in the specific context of Radial Basis Function networks (Moody and Darken, 1989; Poggio and Girosi, 1990; Powell, 1992). Specifically, we are interested in asking the following questions about Radial Basis Functions.

Imagine you were interested in solving a particular problem (regression or pattern classification) using Radial Basis Function networks. Then, how large must the network be and how many examples do you need to draw so that you are guaranteed with high confidence to do very well? Conversely, if you had a finite network and a finite amount of data, what are the kinds of problems you could solve effectively?

Clearly, if one were using a network with a finite number of parameters, then its representational capacity would be limited and therefore even in the best case we would make an approximation error. Drawing upon results in approximation theory (Lorentz, 1986) several researchers (Cybenko, 1989; Barron, 1993; Hornik, Stinchcombe, and White, 1989; Mhaskar, and Micchelli, 1992; Mhaskar, 1993) have investigated the approximating power of feedforward networks showing how as the number of parameters goes to infinity, the network can approximate any continuous function. These results ignore the question of learnability from finite data.

For a finite network, due to finiteness of the data, we make an error in estimating the parameters and consequently have an estimation error in addition to the approximation error mentioned earlier. Using results from Vapnik and Chervonenkis (Vapnik, 1982; Vapnik and Chervonenkis, 1971) and Pollard (Pollard, 1984) , work has also been done (Haussler, 1989; Baum and Haussler, 1989) on the sample complexity of finite networks showing how as the data goes to infinity, the estimation error goes to zero i.e., the empirically optimized parameter settings converge to the optimal ones for that class. However, since the number of parameters are fixed and finite, even the optimal parameter setting might yield a function which is far from the target. This issue is left unexplored by Haussler (1989) in an excellent investigation of the sample complexity question.

In this article we explore the errors due to both finite parameters and finite data in a common setting. In order for the total generalization error to go to zero, both the number of parameters and the number of data have to go to infinity, and we provide rates at which they grow for learnability to result. Further, as a corollary, we are able to provide a principled way of choosing the optimal number of parameters so as to minimize expected errors. It should be mentioned here that A. Barron (1994) and H. White (1990) have also provided treatments of this problem for different hypothesis and concept classes.

The plan of the article is as follows: in section 2 we provide a general formalization of the problem. We then provide in section 3 a precise statement of a specific problem along with our main result, whose proof can be found in (Niyogi, and Girosi, 1994) . In section 4 we discuss what could be the implications of our result in practice; we provide several qualifying remarks and a numerical simulation. Finally we conclude in section 5 with a reiteration of our essential points.

2 Definitions and Statement of the Problem

In order to make a precise statement of the problem we first need to introduce some terminology and to define a number of mathematical objects.

2.1 Random Variables and Probability Distributions

Let X and Y be two arbitrary sets. We will call $\mathbf{x}$ and y the *independent variable* and *response* respectively, where $\mathbf{x}$ and y range over the generic elements of X and Y. In most cases X will be a subset of a k-dimensional Euclidean space and Y a subset of the real line. We assume that an unknown probability distribution $P(\mathbf{x}, y)$ is defined on $X \times Y$.

The probability distribution $P(\mathbf{x}, y)$ can also be written as $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$, where $P(y|\mathbf{x})$ is the conditional probability of the response y given the independent variable $\mathbf{x}$, and $P(\mathbf{x})$ is the marginal probability of the independent variable. Expected values with respect to $P(\mathbf{x}, y)$ or $P(\mathbf{x})$ will be always indicated by $E[\cdot]$. In practical cases we are provided with *examples* of this probabilistic relationship, that is with a data set $D_l \equiv \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^{l}$, obtained by sampling l times the set $X \times Y$ according to $P(\mathbf{x}, y)$. From the definition of $P(\mathbf{x}, y)$ we see that we can think of an element $(\mathbf{x}_i, y_i)$ of the data set D_l as obtained by sampling X according to $P(\mathbf{x})$, and then sampling Y according to $P(y|\mathbf{x})$. The interesting problem is, given an instance of $\mathbf{x}$ that does not appear in the data set D_l, to give an estimate of what we expect y to be.

Formally, we define an *estimator* to be any function $f : X \to Y$. Clearly, since the independent variable $\mathbf{x}$ need not determine uniquely the response y, any estimator will make a certain amount of error. However, it is interesting to study the problem of finding the best possible estimator, given the knowledge of the data set D_l, and this problem will be defined as the problem of *learning from examples*, where the examples are represented by the data set D_l.

2.2 The Expected Risk and the Regression Function

Having defined an estimator, we now need to define a measure of how good an estimator is. Suppose we sample $X \times Y$ according to $P(\mathbf{x}, y)$, obtaining the pair $(\mathbf{x}, y)$. A possible measure of the error of the estimator f at the point $\mathbf{x}$ is $(y - f(\mathbf{x}))^2$. The average error of the estimator f is now given by the functional

$$I[f] \equiv E[(y - f(\mathbf{x}))^2] = \int_{X \times Y} d\mathbf{x}dy \; P(\mathbf{x}, y)(y - f(\mathbf{x}))^2 \ ,$$

that is usually called the *expected risk* of f for the specific choice of the error measure. We are now interested in finding the estimator that minimizes the expected risk over some domain $\mathcal{F}$. We will assume in the following that $\mathcal{F}$ is some space of differentiable functions, for example the space of functions with m bounded derivatives.

Assuming that the problem of minimizing $I[f]$ in $\mathcal{F}$ is well posed, it is easy to obtain its solution. In fact, the expected risk can be decomposed in the following way (see appendix A):

$$I[f] = E[(f_0(\mathbf{x}) - f(\mathbf{x}))^2] + E[(y - f_0(\mathbf{x}))^2] \tag{1}$$

where $f_0(\mathbf{x})$ is the so called *regression function*, that is the conditional mean of the response given the independent variable:

$$f_0(\mathbf{x}) \equiv \int_Y dy \; yP(y|\mathbf{x}) \ . \tag{2}$$

From eq. (1) it is clear that the regression function is the function that minimizes the expected risk in $\mathcal{F}$, and is therefore the best possible estimator. Hence,

$$f_0(\mathbf{x}) = \arg \min_{f \in \mathcal{F}} I[f] \ .$$

While the first term in eq. (1) depends on the choice of the estimator f, the second term is an intrinsic limitation due to the probabilistic nature of the problem, and therefore even the regression function will make an error equal to $E[(y - f_0(\mathbf{x}))^2]$.

The problem of learning from examples can now be reformulated as the problem of reconstructing the regression function f_0, given the example set D_l. It should also be pointed out that this framework includes pattern classification and in this case the regression (target) function corresponds to the Bayes discriminant function (Gish, 1990; Hampshire and Pearlmutter, 1990; Richard and Lippman, 1991).

2.3 The Empirical Risk

In practice the expected risk $I[f]$ is unknown because $P(\mathbf{x}, y)$ is unknown, and our only source of information is the data set D_l. Using this data set, the expected risk can be approximated by the *empirical risk* I_{emp}:

$$I_{\mathrm{emp}}[f] \equiv \frac{1}{l}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))^2 \ .$$

For each given estimator f, the empirical risk is a random variable, and under fairly general assumptions, by the law of large numbers, it converges in probability to the expected risk as the number of data points goes to infinity:

$$\lim_{l\to\infty} P\{|I[f] - I_{\mathrm{emp}}[f]| > \varepsilon\} = 0 \quad \forall \varepsilon > 0 \ . \tag{3}$$

Therefore a common strategy consists in estimating the regression function as the function that minimizes the empirical risk, since it is "close" to the expected risk if the number of data is high enough. However, eq. (3) states only that the expected risk is "close" to the empirical risk *for each given* f, and not for all f *simultaneously.* Consequently the fact that the empirical risk converges in probability to the expected risk when the number, l, of data points goes to infinity does not guarantee that the minimum of the empirical risk will converge to the minimum of the expected risk (the regression function). As pointed out and analyzed in the fundamental work of Vapnik and Chervonenkis (1971, 1991) the notion of *uniform convergence* in probability has to be introduced, and it will be discussed in other parts of this paper.

2.4 The Problem

The argument of the previous section suggests that an approximate solution of the learning problem consists in finding the minimum of the empirical risk, that is solving

$$\min_{f \in \mathcal{F}} I_{\mathrm{emp}}[f] \ .$$

However this problem is often ill-posed, because, for most choices of $\mathcal{F}$, it will have an infinite number of solutions. In fact, all the functions in $\mathcal{F}$ that interpolate the data points $(\mathbf{x}_i, y_i)$, that is with the property $\{f(\mathbf{x}_i) = y_i \quad 1, \ldots, l\}$ will give a zero value for I_{emp}. This problem is very common in approximation/regression theory and statistics and can be approached in several ways. A common technique consists in restricting the search for the minimum to a smaller set than $\mathcal{F}$. We consider the case in which this smaller set is a family of *parametric functions*, that is a family of functions defined by a certain number of real parameters. The choice of a parametric representation also provides a convenient way to store and manipulate the hypothesis function on a computer.

We will denote a generic subset of $\mathcal{F}$ whose elements are parametrized by a number of parameters proportional to n, by H_n. Moreover, we will assume that the sets H_n form a nested family, that is $H_1 \subset H_2 \subset \ldots \subset H_n \subset \ldots \subset H$. For example, H_n could be the set of polynomials in one variable of degree $n-1$, Radial Basis Functions with n centers, Multilayer Perceptrons with n sigmoidal hidden units. Therefore, we choose as approximation to the regression function the function $\hat{f}_{n,l}$ defined as:

$$\hat{f}_{n,l} \equiv \arg \min_{f \in H_n} I_{\mathrm{emp}}[f] \ . \tag{4}$$

It should be pointed out that the sets H_n and $\mathcal{F}$ have to be matched with each other. One could look at this matching from both directions. For a class $\mathcal{F}$, one might be interested in an appropriate choice of H_n. Conversely, for a particular choice of H_n, one might ask what classes $\mathcal{F}$ can be effectively solved with this scheme.

Thus, we see that in principle we would like to minimize $I[f]$ over the large class $\mathcal{F}$ obtaining thereby the regression function f_0. What we do in practice is to minimize the empirical risk $I_{\mathrm{emp}}[f]$ over the smaller class H_n obtaining the function $\hat{f}_{n,l}$. Assuming we have solved all the computational problems related to the actual computation of the estimator $\hat{f}_{n,l}$, the main problem is now: **how good is $\hat{\mathbf{f}}_{\mathbf{n,l}}$?**
Independently of the measure of performance that we choose when answering this question, we expect $\hat{f}_{n,l}$ to become a better and better estimator as n and l go to infinity. In fact,

when l increases, our estimate of the expected risk improves and our estimator improves. The case of n is trickier. As n increases, we have more parameters to model the regression function, and our estimator should improve. However, at the same time, because we have more parameters to estimate with the same amount of data, our estimate of the expected risk deteriorates. Thus we now need more data and n and l have to grow as a function of each other for convergence to occur. At what rate and under what conditions the estimator $\hat{f}_{n,l}$ improves depends on the properties of the regression function, that is on $\mathcal{F}$, and on the approximation scheme we are using, that is on H_n.

2.5 Bounding the Generalization Error

Recall that our goal is to minimize the expected risk $I[f]$ over the set $\mathcal{F}$. If instead we were to choose our estimator from H_n we would obtain f_n as:

$$f_n \equiv \arg \min_{f \in H_n} I[f] \ .$$

However we can only minimize the empirical risk I_{emp}, obtaining as our real estimate the function $\hat{f}_{n,l}$. Our goal is to bound the distance from $\hat{f}_{n,l}$ to f_0. If we choose to measure the distance in the $L^2(P)$ metric, the quantity that we need to bound, that we will call *generalization error*, is:

$$E[(f_0 - \hat{f}_{n,l})^2] = \int_X d\mathbf{x}\ P(\mathbf{x})(f_0(\mathbf{x}) - \hat{f}_{n,l}(\mathbf{x}))^2 = \|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)}$$

There are 2 main factors that contribute to the generalization error, and we are going to analyze them separately for the moment.
1. A first source of error is due to the fact that we are trying to approximate an infinite dimensional object, the regression function $f_0 \in \mathcal{F}$, with a function defined by a finite number of parameters. We call this *the approximation error*, and we measure it by the quantity $E[(f_0 - f_n)^2]$. The approximation error can be expressed in terms of the expected risk using the decomposition (1) as

$$E[(f_0 - f_n)^2] = I[f_n] - I[f_0] \ . \tag{5}$$

Notice that the approximation error does not depend on the data set D_l, but depends only on the approximating power of the class H_n, and can be naturally studied within the framework of approximation theory. In the following we will always assume that it is possible to bound the approximation error as follows:

$$E[(f_0 - f_n)^2] \leq \varepsilon(n)$$

where $\varepsilon(n)$ is a function that goes to zero as n goes to infinity if H is dense in $\mathcal{F}$. In other words, as the number n of parameters gets larger the representation capacity of H_n increases, and allows a better and better approximation of the regression function f_0. This issue has been studied by a number of researchers (Cybenko, 1989; Hornik, Stinchcombe and White, 1989; Jones, 1992; Barron, 1993; Mhaskar and Micchelli, 1992; Mhaskar, 1993) in the neural networks community.

2. Another source of error comes from the fact that, due to finite data, we minimize the empirical risk $I_{\text{emp}}[f]$, and obtain $\hat{f}_{n,l}$, rather than minimizing the expected risk $I[f]$, and obtaining f_n. As the number of data goes to infinity we hope that $\hat{f}_{n,l}$ will converge to f_n, and convergence will take place if the empirical risk converges to the expected risk *uniformly in probability* (Vapnik, 1982). The quantity $|I_{\text{emp}}[f] - I[f]|$ is called *estimation error*, and conditions for the estimation error to converge to zero uniformly in probability have been investigated by Vapnik and Chervonenkis (1971,1991), Pollard (1984) , Dudley (1987), and Haussler (1989) . Under a variety of different hypothesis it is possible to prove that, with probability $1 - \delta$, a bound of this form is valid:

$$|I_{\text{emp}}[f] - I[f]| \leq \omega(l, n, \delta) \quad \forall f \in H_n \tag{6}$$

The specific form of ω depends on the setting of the problem, but, in general, we expect $\omega(l, n, \delta)$ to be a decreasing function of l. However, we also expect it to be an increasing function of n. The reason is that, if the number of parameters is large then the expected risk is a very complex object, and then more data will be needed to estimate it. Therefore, keeping fixed the number of data and increasing the number of parameters will result, on the average, in a larger distance between the expected risk and the empirical risk.

The approximation and estimation error are clearly two components of the generalization error, and it is interesting to notice, as shown in the next statement and represented in figure (1), the generalization error can be bounded by a linear combination of the two:

Statement 2.1 *The following inequality holds:*

$$\|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)} \leq \varepsilon(n) + 2\omega(l, n, \delta) \ . \tag{7}$$

FIGURE (1) HERE

Proof: using the decomposition of the expected risk (1), the generalization error can be written as:

$$\|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)} = E[(f_0 - \hat{f}_{n,l})^2] = I[\hat{f}_{n,l}] - I[f_0] \ . \tag{8}$$

A natural way of bounding the generalization error is as follows:

$$E[(f_0 - \hat{f}_{n,l})^2] \leq |I[f_n] - I[f_0]| + |I[f_n] - I[\hat{f}_{n,l}]| \ . \tag{9}$$

In the first term of the right hand side of the previous inequality we recognize the approximation error (5). If a bound of the form (6) is known for the estimation error, it is simple to show (see figure 2) that the second term can be bounded as

$$|I[f_n] - I[\hat{f}_{n,l}]| \leq 2\omega(l, n, \delta)$$

and statement (2.1) follows $\Box$.

FIGURE (2) HERE

A Note on Models and Model Complexity: from the form of eq. (7) the reader will realize that there is a trade-off between n and l for a certain generalization error. For a fixed l, as n increases, the approximation error $\varepsilon(n)$ decreases but the estimation error $\omega(l, n, \delta)$ increases. Consequently, there is a certain n which might optimally balance this trade-off. Note that the classes H_n can be looked upon as models of increasing complexity and the search for an optimal n amounts to a search for the right model complexity. One typically wishes to match the model complexity with the sample complexity (measured by how much data we have on hand) and this problem is well studied (Rissanen, 1989; Barron and Cover, 1989; Efron, 1982; Craven and Wahba, 1979) in statistics. Broadly speaking, simple models would have high approximation errors but small estimation errors while complex models would have low approximation errors but high estimation errors. This trade-off is also embodied in the so-called bias-variance dilemma as described in Geman et al. (1992).

So far we have provided a very general characterization of this problem, without stating what the sets $\mathcal{F}$ and H_n are, and in the next section we will consider a specific choice for these sets, and we will provide a bound on the generalization error of the form of eq. (7).

3 Stating the Problem for Radial Basis Functions

In this article we focus our attention on a Radial Basis Functions approximation scheme . This is an *hypothesis* class defined as follows:

$$H_n \equiv \left\{ f \mid f(\mathbf{x}) = \sum_{i=1}^{n} \beta_i G\left(\frac{\|\mathbf{x} - \mathbf{t}_i\|}{\sigma_i}\right) \right\} \tag{10}$$

where G is a Gaussian function and the β_i, $\mathbf{t}_i$, and σ_i are free parameters. We would like to understand what classes of problems can be solved "well" by this technique, where "well" means that both approximation and estimation bounds need to be favorable. It is possible to show that a favorable approximation bound can be obtained if we assume that the *concept* class of functions $\mathcal{F}$ to which the regression function belongs is defined as follows:

$$\mathcal{F} \equiv \{f | f = \lambda * G_m, m > k/2, |\lambda|_{R^k} \leq M\} \ . \tag{11}$$

Here M is a positive number, λ is a signed Radon measure on the Borel sets of R^k and G_m is the Bessel-Macdonald kernel, i.e., the inverse fourier transform of $\tilde{G}_m(\mathbf{s}) = (1 + \|\mathbf{s}\|^2)^{-m/2}$. The symbol $*$ stands for the convolution operation, $|\lambda|_{R^k}$ is the total variation of the measure λ. The space $\mathcal{F}$ as defined in eq. 11 is the *Bessel potential space* of order m, $\mathcal{L}_1^m$. If m is even, this contains the *Sobolev Space* $H^{m,1}$ of functions whose derivatives upto order m are integrable (Stein, 1970).

In order to obtain an estimation bound we need the approximating class to have bounded variation, and we impose the constraint $\sum_{i=1}^{n} |\beta_i| \leq M$. This constraint does not affect the approximation bound, and the two pieces fit together nicely. Thus the set H_n is defined now as the set of functions belonging to L_2 such that

$$f(\mathbf{x}) = \sum_{i=1}^{n} \beta_i G\left(\frac{\|\mathbf{x} - \mathbf{t}_i\|}{\sigma_i}\right) \ , \ \sum_{i=1}^{n} |\beta_i| \leq M \ , \quad \mathbf{t}_i \in R^k \ , \ \sigma_i \in R^+ \ , \forall i = 1, \ldots, n \ . \tag{12}$$

Having defined the sets H_n and $\mathcal{F}$ we remind the reader that our goal is to recover the regression function, that is the minimum of the expected risk over $\mathcal{F}$. What we end up doing is to draw a set of l examples and to minimize the empirical risk I_{emp} over the set H_n, that is to solve the following non-convex minimization problem:

$$\hat{f}_{n,l} \equiv \arg \min_{\beta_\alpha, \mathbf{t}_\alpha, \sigma_\alpha} \sum_{i=1}^{l} \left(y_i - \sum_{\alpha=1}^{n} \beta_\alpha G \left(\frac{\|\mathbf{x}_i - \mathbf{t}_\alpha\|}{\sigma_\alpha} \right) \right)^2 \tag{13}$$

Assuming now that we have been able to solve the minimization problem of eq. (13), the main question we are interested in is "how far is $\hat{f}_{n,l}$ from f_0?". We give an answer in the next section.

3.1 Main Result

Our main theorem is now stated in a PAC-like formulation:

Theorem 3.1 *Let H_n be the class of Gaussian RBF networks with k input nodes and n hidden nodes as defined in eq. 10, and f_0 be an element of the Bessel potential space $\mathcal{L}_1^m(R^k)$ of order m, with $m > k/2$ (the class $\mathcal{F}$ defined in eq. 11). Assume that a data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ has been obtained by randomly sampling the function f_0 in presence of noise, and that the noise distribution has compact support. Then, for any $0 < \delta < 1$, with probability greater than $1 - \delta$, the following bound for the generalization error holds:*

$$\|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)} \leq O\left(\frac{1}{n}\right) + O\left(\left[\frac{nk\ln(nl) - \ln\delta}{l}\right]^{1/2}\right) \tag{14}$$

This theorem is proved by decomposing the total generalization error into an approximation component and an estimation one as in eq. 7. The bound for the approximation error (the first term in the equation above) can be found in (Girosi, 1994) and (Girosi and Anzellotti, 1993) , and it is a consequence of the Maurey-Jones-Barron lemma (Jones, 1992; Barron, 1993). The bound for the estimation error (the second term) has been obtained using ideas from the uniform convergence of empirical estimates to their means (Vapnik, 1982). In particular, we have used notions of metric entropy (Pollard, 1984) to bound the complexity of the class H_n. The full proof of this theorem is not reported here because of its length, and can be found in (Niyogi and Girosi, 1994).

4 Implications of the Theorem in Practice: Putting In the Numbers

In figure (3) we show the bound on the generalization error presented in the previous section as a function of the number of examples (l) and the number of basis functions (n). A number of remarks about this picture are in order.

FIGURE (3) HERE

4.1 Rate of Growth of n for Guaranteed Convergence

From our theorem (3.1) we see that the generalization error converges to zero only if n goes to infinity more slowly than l. In fact, if n grows too quickly the estimation error $\omega(l, n, \delta)$ will diverge, because it is proportional to n. In fact, setting $n = l^r$, we obtain

$$\lim_{l \to +\infty} \omega(l, n, \delta) = \lim_{l \to +\infty} l^{r-1} \ln l \ .$$

Therefore the condition $r < 1$ should hold in order to guarantee convergence to zero.

4.2 Optimal Choice of n

In the previous section we made the point that the number of parameters n should grow more slowly than the number of data points l, in order to guarantee the consistency of the estimator $\hat{f}_{n,l}$. It is quite clear that there is an *optimal* rate of growth of the number of parameters, that, for any fixed amount of data points l, gives the best possible performance with the least number of parameters. In other words, for any fixed l there is an optimal number of parameters $n^*(l)$ that minimizes the generalization error. That such a number should exist is quite intuitive: for a fixed number of data, a small number of parameters will give a low estimation error $\omega(l, n, \delta)$, but very high approximation error $\varepsilon(n)$, and therefore the generalization error will be high. If the number of parameters is very high the approximation error $\varepsilon(n)$ will be very small, but the estimation error $\omega(l, n, \delta)$ will be high, leading to a large generalization error again. Therefore, somewhere in between there should be a number of parameters high enough to make the approximation error small, but not too high, so that these parameters can be estimated reliably, with a small estimation error. Although the exact form for the generalization error is unknown, we can work with the upper bound (14), which we plot in figure (4) as a function of the number of parameters n for various choices of sample size l. Notice that for a fixed sample size, the error passes through a minimum. Notice that the location of the minimum shifts to the right when the sample size is increased.

FIGURE (4) HERE

In order to find out exactly what is the optimal rate of growth of the network size we simply find the minimum of the generalization error as a function of n keeping the sample size l fixed. Therefore we have to solve the equation:

$$\frac{\partial}{\partial n} E[(f_0 - \hat{f}_{n,l})^2] = 0$$

for n as a function of l. Substituting the bound given in theorem (3.1) in the previous equation, and ignoring logarithmic factors, we obtain an approximation of the optimal number of parameters $n^*(l)$ for a given number of examples l behaves as

$$n^*(l) \propto l^{\frac{1}{3}} \ . \tag{15}$$

While a fixed sample size suggests the scheme above for choosing an optimal network size, it is important to note that for a certain confidence rate (δ) and for a fixed error bound, there are various choices of n and l which are satisfactory. Fig. 5 shows n as a function of l, in other words (n, l) pairs which yield the same error bound (E) with the same confidence.

FIGURE (5) HERE

For any fixed error bound, the region to the right of the minimum is uninteresting because it uses more parameters and data than needed. The narrow region between the minimum and the asymptote is more interesting: if networks size is very expensive, less parameters can be used at the expenses of many more data points. Notice however how narrow this region is and how quickly the curve goes to infinity: a very large number of data points is needed to compensate for a little less parameters.

4.3 Remarks

In this section we suggest future work, and make connections with other related research.
Extensions:

1. While we have obtained an upper bound on the error in terms of the number of nodes and examples, it would be worthwhile to obtain lower bounds on the same. Such lower bounds do not seem to exist in the neural network literature to the best of our knowledge.
2. We have considered here a situation where the estimated network i.e., $\hat{f}_{n,l}$ is obtained by minimizing the empirical risk over the class of functions H_n. Very often, the estimated network is obtained by minimizing a somewhat different objective function which consists of two parts. One is the fit to the data and the other is some complexity term which favors less complex (according to the defined notion of complexity) functions over more complex ones. For example the regularization approach (Tikhonov, 1963; Poggio, and Girosi, 1990; Wahba, 1990) minimizes a cost function of the form

$$H[f] = \sum_{i=1}^{N}(y_i - f(\mathbf{x}_i) + \lambda\Phi[f]$$

over the class $H = \cup_{n\geq 1} H_n$. Here λ is the so called "regularization parameter" and $\Phi[f]$ is a functional which measures smoothness of the functions involved. Choice of an optimal λ is an interesting question in regularization techniques and typically cross-validation or other heuristic schemes are used.

3. Structural risk minimization (Vapnik, 1982) is another method to achieve a trade-off between network complexity (corresponding to n in our case) and fit to data. However it does not guarantee that the architecture selected will be the one with minimal parametrization. In fact, it would be of some interest to develop a sequential growing scheme. Such a technique would at any stage perform a sequential hypothesis test. It would then decide whether to ask for more data, add one more node or simply stop and output the function it has as its ϵ-good hypothesis. In such a process, one might even incorporate active learning (Angluin, 1988; Niyogi, 1995) so that if the algorithm asks for more data, then it might even specify a region in the input domain from where it would like to see this data.
4. It should be noted here that we have assumed that the empirical risk $\sum_{i=1}^{l}(y_i - f(x_i))^2$ can be minimized over the class H_n and the function $\hat{f}_{n,l}$ be effectively computed. While this might be fine in principle, in practice only a locally optimal solution to the minimization problem is found (typically using some gradient descent schemes). The computational complexity of obtaining even an approximate solution to the minimization problem is an

interesting one and results from computer science (Judd, 1988; Blum, and Rivest, 1988) suggest that it might in general be NP-hard.

Connections with Other Results

1. In the neural network and computational learning theory communities results have been obtained pertaining to the issues of generalization and learnability. Some theoretical work has been done (Baum and Haussler, 1989; Haussler, 1989; Ji and Psaltis, 1992) in characterizing the sample complexity of finite sized networks. Of these, it is worthwhile to mention again the work of Haussler (1989) from which this paper derives much inspiration. He obtains bounds for a fixed hypothesis space i.e. a fixed finite network architecture. Here we deal with families of hypothesis spaces using richer and richer hypothesis spaces as more and more data becomes available. Others (Levin, Tishby and Solla, 1990) attempt to characterize the generalization abilities of feed-forward networks using theoretical formalizations from statistical mechanics. Yet others (Botros and Atkeson, 1991; Moody, 1992; Cohn and Tesauro, 1991; Rumelhart, Weigand, and Huberman, 1991) attempt to obtain empirical bounds on generalization abilities.

2. This is an attempt to obtain rate-of-convergence bounds in the spirit of Barron's work (1994), but using a different approach. We have chosen to combine theorems from approximation theory (which gives us the $O(1/n)$ term in the rate, and uniform convergence theory (which gives us the other part). Note that at this moment, our rate of convergence is worse than Barron's. In particular, he obtains a rate of convergence of $O(1/n + (nk\ln(l))/l)$. Further, he has a different set of assumptions on the class of functions (corresponding to our $\mathcal{F}$). Finally, the approximation scheme is a class of networks with sigmoidal units as opposed to radial-basis units and a different proof technique is used.

3. It is worthwhile to refer to the article of Geman, Bienenstock, and Doursat (1992) in this journal which discusses the Bias-Variance dilemma. Using our notation the integrated square bias is defined as $B = \|f_0 - E_{D_l}[\hat{f}_{n,l}]\|^2$ and the integrated variance is $V = E_{D_l}[(E_{D_l}[\hat{f}_{n,l}] - \hat{f}_{n,l})^2]$, where E_{D_l} stands for the expected value over all the possible data sets of size l. Geman, Bienenstock, and Doursat (1992) show that the generalization error averaged over D_l can be decomposed as $B + V$. They show that as the number of parameters increases, the bias of the estimator decreases and the variance increases for a fixed size of the data set. From an intuitive point of view, the bias B plays the role of the approximation error $\|f_0 - f_n\|^2$, although their relationship is not clear. In fact, the average estimator $E_{D_l}[\hat{f}_{n,l}]$ differs from f_n, and need not even belong to H_n. The variance V is related to the average estimation error, and it can be shown that both of them are bounded by the quantity $E_{D_l}\|f_n - \hat{f}_{n,l}\|^2$. Finding the right bias-variance trade-off is very

similar in spirit to finding the trade-off between network complexity and data complexity.
4. Given the class of radial basis functions we are using, a natural comparison arises with kernel regression (Krzyzak, 1986; Devroye, 1981) and results on the convergence of kernel estimators. It should be pointed out that, unlike our scheme, Gaussian-kernel regressors require the variance of the Gaussian to go to zero as a function of the data. Further the number of kernels is always equal to the number of data points and the issue of trade-off between the two is not explored to the same degree.
5. In our statement of the problem, we discussed how pattern classification could be treated as a special case of regression. In this case the function f_0 corresponds to the Bayes *a-posteriori* decision function. Researchers (Richard, and Lippman, 1991; Hampshire, and Pearlmutter, 1990; Gish, 1990) in the neural network community have observed that a network trained on a least square error criterion and used for pattern classification was in effect computing the Bayes decision function. This paper provides a rigorous proof of the conditions under which this is the case.

4.4 Empirical Results

The main thrust of this paper is to provide some insight into how overfitting can be studied in classes of feedforward networks and the general laws that govern overfitting phenomena in such networks. How closely do "real" function learning problems obey the the general principles embodied in the theorem described earlier? We do not attempt to provide an extensive answer to this question—but just to satisfy the reader's curiosity, we now describe some empirical results.

The experiment: The target function, a k-dimensional function, was assumed to have the following form, which ensures that the assumptions of theorem (3.1) are satisfied:

$$f(\mathbf{x}) = \left(\sum_{i=1}^{N} c_i \sin(\mathbf{x} \cdot \mathbf{w}_i + \phi_i) \right) e^{-\|\Sigma^{-1}\mathbf{x}\|^2} \tag{16}$$

Here Σ is a diagonal matrix $(\Sigma)_{\alpha\beta} = k\ \sigma_\alpha \delta_{\alpha\beta}$. The parameters, $\{\sigma_\alpha, \mathbf{w}_i, c_i\}$ were chosen at random in the following ranges: $\sigma_i \in [1.7, 2.3]$, $\mathbf{w}_i \in [-2, 2]^k$, $c_i \in [-1, 1]$, $\phi_i \in [0, 2\pi]$, $N \in [3, 20]$. Training sets of different sizes, ranging from $l = 30$ to $l = 500$ were randomly generated in the k dimensional cube $[-\pi, \pi]^k$, and an independent test set of 2000 examples was chosen to estimate the generalization error. Gaussian RBF networks (as in theorem 3.1) with different number of hidden units, ranging from $n = 1$ to $n = 300$, were trained using a gradient descent scheme. Each training session was repeated 10 times with random initialization, because of the problem of local minima. We did experiments in 2, 4, 6 and 8 dimensions. In all cases the qualitative behavior of the experimental results followed the theoretical predictions. In figures 6 and 7 we report the experimental results for a 2 and 6 dimensional case respectively.

FIGURE (6) AND (7) HERE

We found, in general, that although overfitting occurs as expected, it has a tendency to occurr at a larger number of parameters than expected. We attribute that to the presence of local minima, that have the effect of restricting the hypothesis, and suggesting that the "effective" number of parameters (Moody, 1992) is much smaller than the total number of parameters.

We believe that extensive experimentation is needed to compare the deviation between theory and practice, and the problem of local minima should be seriously addressed. This is well beyond the scope of the current article, and further research on the matter is planned.

5 Conclusion

For the task of learning some unknown function from labeled examples where we have multiple hypothesis classes of varying complexity, choosing the class of right complexity and the appropriate hypothesis within that class poses an interesting problem. We have provided an analysis of the situation and the issues involved and in particular have tried to show how the hypothesis complexity, the sample complexity and the generalization error are related. We proved a theorem for a special set of hypothesis classes, the radial basis function networks, and we bound the generalization error for certain function learning tasks in terms of the number of parameters and the number of examples. This is equivalent to obtaining a bound on the rate at which the number of parameters must grow with respect to the number of examples for convergence to take place. Thus we use richer and richer hypothesis spaces as more and more data become available. We also see that there is a tradeoff between hypothesis complexity and generalization error for a certain fixed amount of data and our result allows us a principled way of choosing an appropriate hypothesis complexity (network architecture). The choice of an appropriate model for empirical data is a problem of long-standing interest in statistics and we provide connections between our work and other work in the field.

Acknowledgments We are grateful to V. Vapnik, T. Poggio and B. Caprile for useful discussions and suggestions. We also wish to thank N.T. Chan for kindly providing the code for the numerical simulations.

A A Useful Decomposition of the Expected Risk

We now show that regression function defined in eq. (2) minimizes the expected risk, $I[f]$. By adding and subtracting the regression function, f_0, we see that:

$$\begin{aligned} I[f] &= \int_{X\times Y} d\mathbf{x}dy P(\mathbf{x},y)(y - f_0(\mathbf{x}) + f_0(\mathbf{x}) - f(\mathbf{x}))^2 = \\ &= \int_{X\times Y} d\mathbf{x}dy P(\mathbf{x},y)(y - f_0(\mathbf{x}))^2 + \\ &+ \int_{X\times Y} d\mathbf{x}dy P(\mathbf{x},y)(f_0(\mathbf{x}) - f(\mathbf{x}))^2 + \\ &+ 2\int_{X\times Y} d\mathbf{x}dy P(\mathbf{x},y)(y - f_0(\mathbf{x}))(f_0(\mathbf{x}) - f(\mathbf{x})) \end{aligned}$$

By definition of the regression function $f_0(\mathbf{x})$, the cross product in the last equation is easily seen to be zero, and therefore

$$I[f] = \int_X d\mathbf{x} P(\mathbf{x})(f_0(\mathbf{x}) - f(\mathbf{x}))^2 + I[f_0] \; .$$

Clearly, the minimum of $I[f]$ is achieved when the first term is minimum, that is when $f(\mathbf{x}) = f_0(\mathbf{x})$. In the case in which the data come from randomly sampling a function f in presence of additive noise, $I[f_0] = \sigma^2$ where σ^2 is the variance of the noise. When data are noisy, therefore, even in the most favourable case we cannot expect the expected risk to be smaller than the variance of the noise.

References

[1] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

[2] A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transaction on Information Theory*, 39(3):930–945, May 1993.

[3] A. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133, 1994.

[4] A. Barron and T. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4), 1991.

[5] E.B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.

[6] A. Blum and R. L. Rivest. Training a three-neuron neural net is NP-complete. In *Proceedings of the 1988 Workshop on Computational LearningTheory*, pages 9–18, San Mateo, CA, 1988. Morgan Kaufma.

[7] S. Botros and C. G. Atkeson. Generalization properties of Radial Basis Functions. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural information processings systems 3*, pages 707–713, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[8] D. Cohn and G. Tesauro. Can neural networks do better than the VC bounds. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural information processings systems 3*, pages 911–917, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[9] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation. *Numer. Math*, 31:377–403, 1979.

[10] G. Cybenko. Approximation by superposition of a sigmoidal function. *Math. Control Systems Signals*, 2(4):303–314, 1989.

[11] L. Devroye. On the almost everywhere convergence of nonparametric regression function estimate. *The Annals of Statistics*, 9:1310–1319, 1981.

[12] R.M. Dudley. Universal Donsker classes and metric entropy. *Ann. Prob.*, 14(4):1306–1326, 1987.

[13] B. Efron. The jacknife, the bootstrap, and other resampling plans. *SIAM, Philadelphia*, 1982.

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

[15] F. Girosi. Regularization theory, Radial Basis Functions and networks. In V. Cherkassky, J.H. Friedman, and H. Wechsler, editors, *From Statistics to Neural Networks. Theory and Pattern Recognition Applications*. Springer-Verlag, Subseries F, Computer and Systems Sciences, 1994.

[16] F. Girosi and G. Anzellotti. Rates of convergence for Radial Basis Functions and neural networks. In R.J. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 97–113, London, 1993. Chapman & Hall.

[17] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

[18] H. Gish. A probabilistic approach to the understanding and training of neural network classifiers. In *Proceedings of the ICASSP-90*, pages 1361–1365, Albuquerque, New Mexico, 1990.

[19] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. Technical Report UCSC-CRL-91-02, University of California, Santa Cruz, 1989.

[20] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[21] J. B. Hampshire II and B. A. Pearlmutter. Equivalence proofs for multilayer perceptron classifiers and the bayesian discriminant function. In J. Elman D. Touretzky and G. Hinton, editors, *Proceedings of the 1990 Connectionist Models Summer School*, San Mateo, CA, 1990. Morgan Kaufman.

[22] C. Ji and D. Psaltis. The VC dimension versus the statistical capacity of multilayer networks. In S. J. Hanson J. Moody and R. P. Lippman, editors, *Advances in Neural information processings systems 4*, pages 928–935, San Mateo, CA, 1992. Morgan Kaufman.

[23] L.K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for Projection Pursuit Regression and neural network training. *The Annals of Statistics*, 20(1):608–613, March 1992.

[24] S. Judd. *Neural Network Design and the Complexity of Learning*. PhD thesis, University of Massachusetts, Amherst, Amherst, MA, 1988.

[25] A. Krzyzak. The rates of convergence of kernel regression estimates and classification rules. *IEEE Transactions on Information Theory*, IT-32(5):668–679, September 1986.

[26] E. Levin, N. Tishby, and S. A. Solla. A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE*, 78(10):1568–1574, October 1990.

[27] R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, pages 4–22, April 1987.

[28] G. G. Lorentz. *Approximation of Functions*. Chelsea Publishing Co., New York, 1986.

[29] H.N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1:61–80, 1993.

[30] H.N. Mhaskar and C.A. Micchelli. Approximation by superposition of a sigmoidal function. *Advances in Applied Mathematics*, 13:350–373, 1992.

[31] J. Moody. The effective number of parameters: An analysis of generalization and regularization in non-linear learning systems. In S. J. Hanson J. Moody and R. P. Lippman, editors, *Advances in Neural information processings systems 4*, pages 847–854, San Mateo, CA, 1991. Morgan Kaufman.

[32] J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

[33] P. Niyogi. *The informational complexity of learning from examples.* PhD thesis, MIT, February 1995.

[34] P. Niyogi and F. Girosi. On the relationship between generalization error, hypothesis complexity, and sample complexity for Radial Basis Functions. A.I. Memo 1467, Massachussetts Institute of Technology, 1994. URL ftp://publications.ai.mit.edu/ai-publications/1000-1499/AIM-1467.ps.Z.

[35] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), September 1990.

[36] D. Pollard. *Convergence of stochastic processes.* Springer-Verlag, Berlin, 1984.

[37] M.J.D. Powell. The theory of radial basis functions approximation in 1990. In W.A. Light, editor, *Advances in Numerical Analysis Volume II: Wavelets, Subdivision Algorithms and Radial Basis Functions*, pages 105–210. Oxford University Press, 1992.

[38] M. D. Richard and R. P. Lippman. Neural network classifier estimates bayesian a-posteriori probabilities. *Neural Computation*, 3:461–483, 1991.

[39] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, 1989.

[40] E.M. Stein. *Singular integrals and differentiability properties of functions.* Princeton, N.J., Princeton University Press, 1970.

[41] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

[42] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, Berlin, 1982.

[43] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequences of events to their probabilities. *Th. Prob. and its Applications*, 17(2):264–280, 1971.

[44] V.N. Vapnik and A. Y. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):283–305, 1991.

[45] G. Wahba. *Splines Models for Observational Data.* Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.

[46] A. S. Weigand, D. E. Rumelhart, and B. A. Huberman. Generalization by weight elimination with applications to forecasting. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural information processings systems 3*, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[47] H. White. Connectionist nonparametric regression: Multilayer perceptrons can learn arbitrary mappings. *Neural Networks*, 3(535-549), 1990.

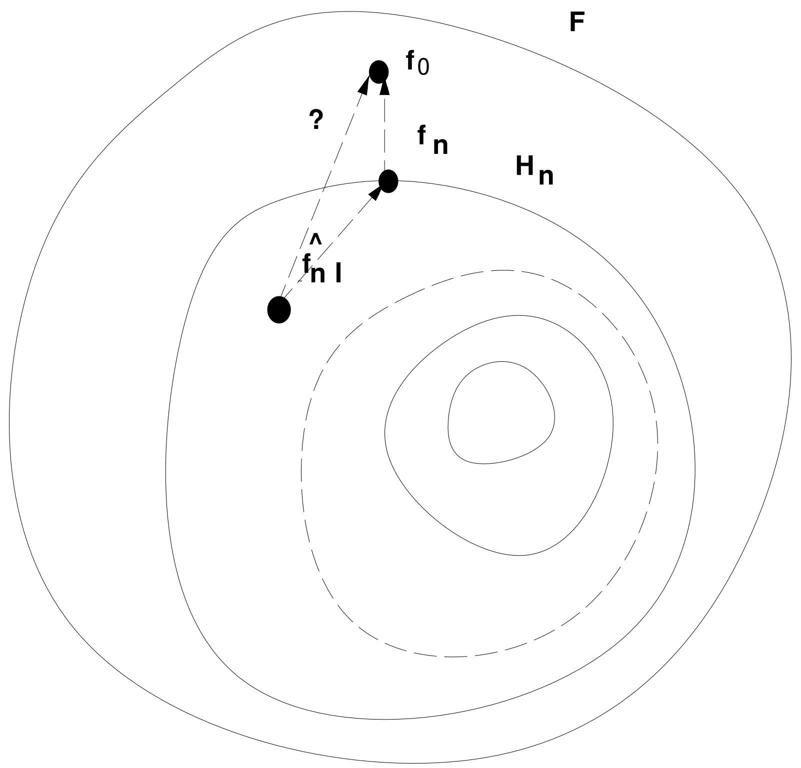

Figure 1: The outermost circle represents the concept class $\mathcal{F}$. Embedded in this are the nested approximating subsets H_n (hypothesis classes). The target function f_0 is an element of $\mathcal{F}$, f_n is the closest element of H_n to f_0, and $\hat{f}_{n,l}$ is the element of H_n which the learner hypothesizes on the basis of data. The arrow with the question mark represents the generalization error, and the other two arrows represent the approximation and estimation errors.

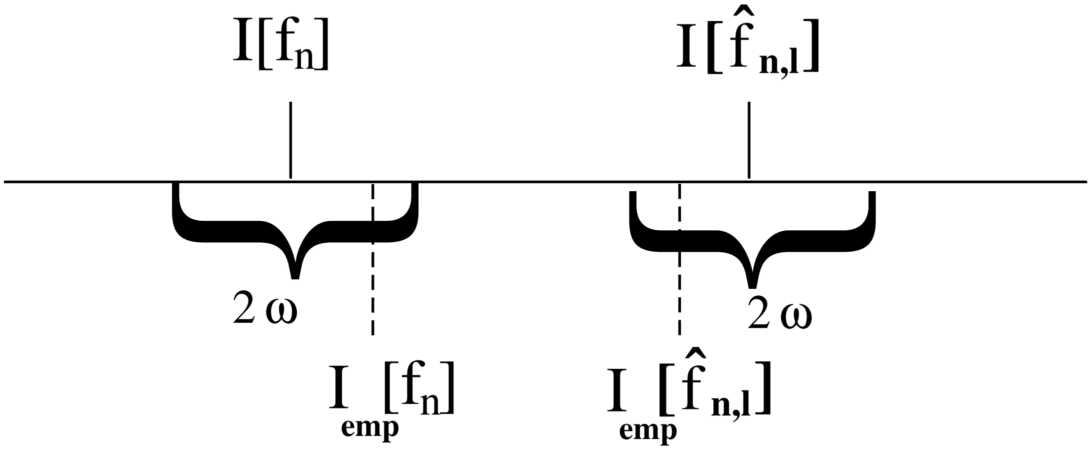

Figure 2: This picture represents the fact that $I[f_n] \leq I[\hat{f}_{n,l}]$ and that $|I[f] - I_{\mathrm{emp}}[f]| \leq \omega$ for all $f \in H_n$. Notice that if the distance between $I[f_n]$ and $I[\hat{f}_{n,l}]$ is larger than 2ω, the condition $I_{\mathrm{emp}}[\hat{f}_{n,l}] \leq I_{\mathrm{emp}}[f_n]$ is violated, and therefore we must have that $|I[f_n] - I[\hat{f}_{n,l}]| \leq 2\omega$.

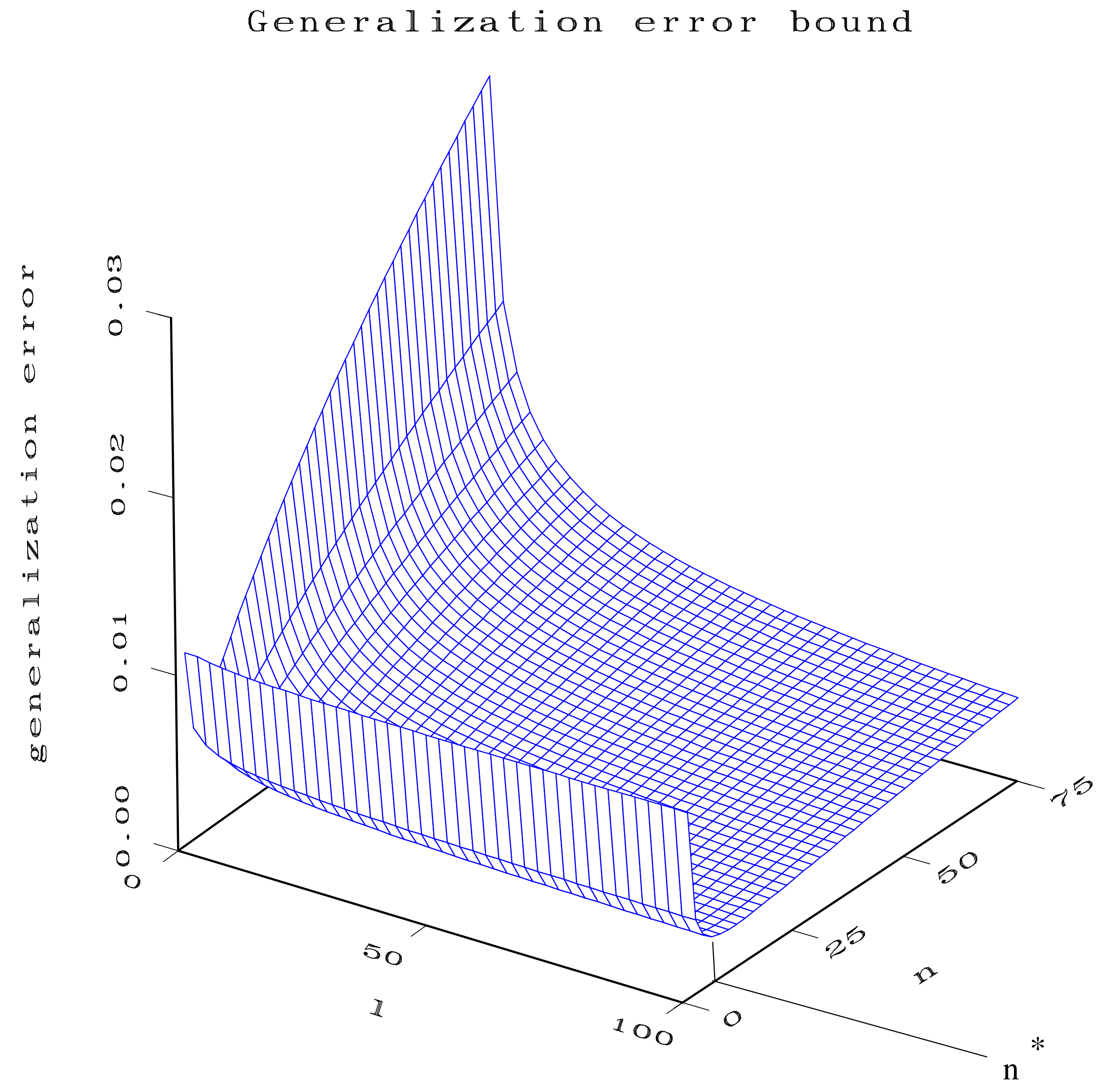

Figure 3: The bound on the generalization error derived in theorem (3.1) plotted as a function of the number of examples (l) and the number of basis functions (n). The bound has the form $\frac{a}{n} + b[(nk\ln(nl) - \ln\delta)/l]^{1/2}$, and in this picture the parameters have values $a = 0.01$, $b = 0.0006$ $k = 5$ and $\delta = 0.01$. For $l = 100$ we show n^*, the critical number of nodes after which overfitting occurs.

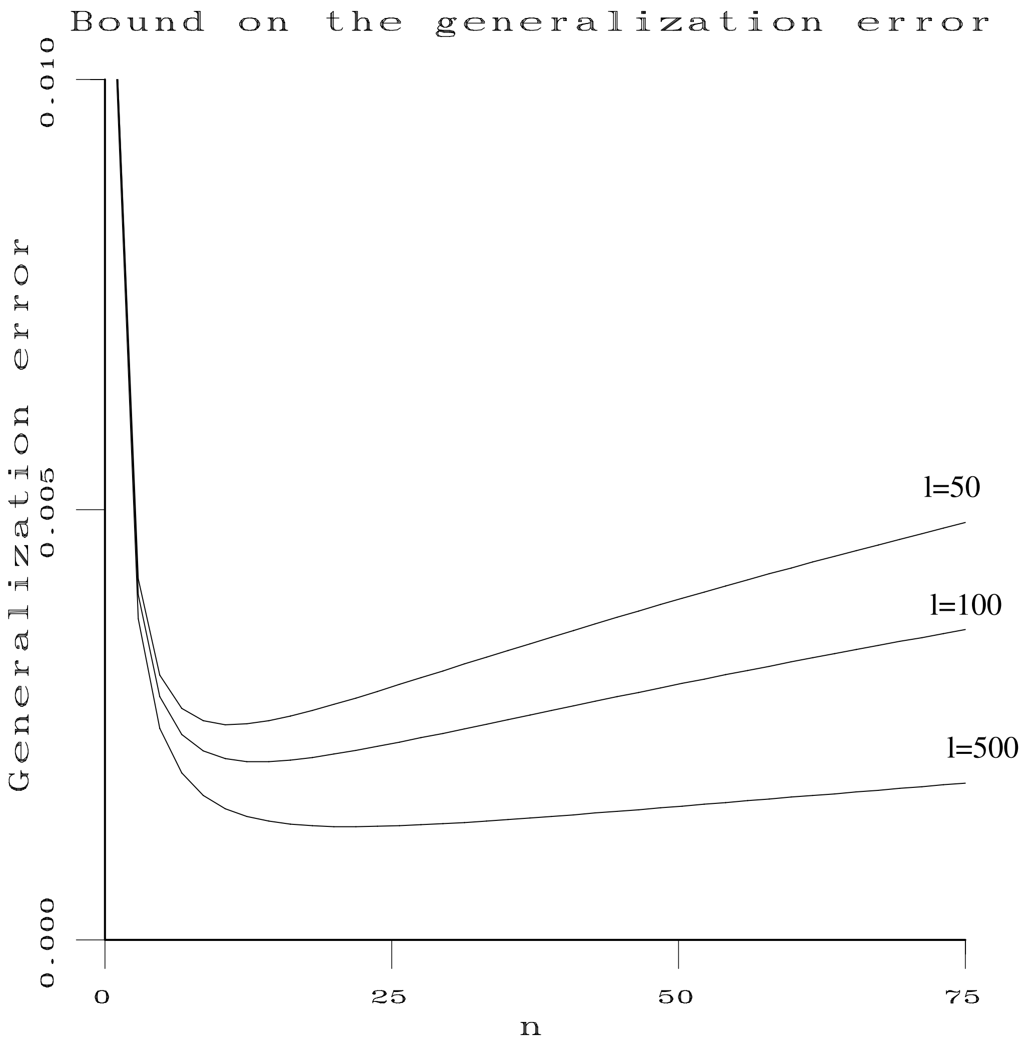

Figure 4: The bound (14) on the generalization error is here plotted as a function of the number of basis functions n, for different number of data points ($l = 50, 100, 500$). The parameters are the same as in figure (3). Notice how the minima $n^*(l)$ of these curves move as l increases. Note also that the minima are broader for larger l, suggesting that an accurate choice of n is less critical when plenty of data is available.

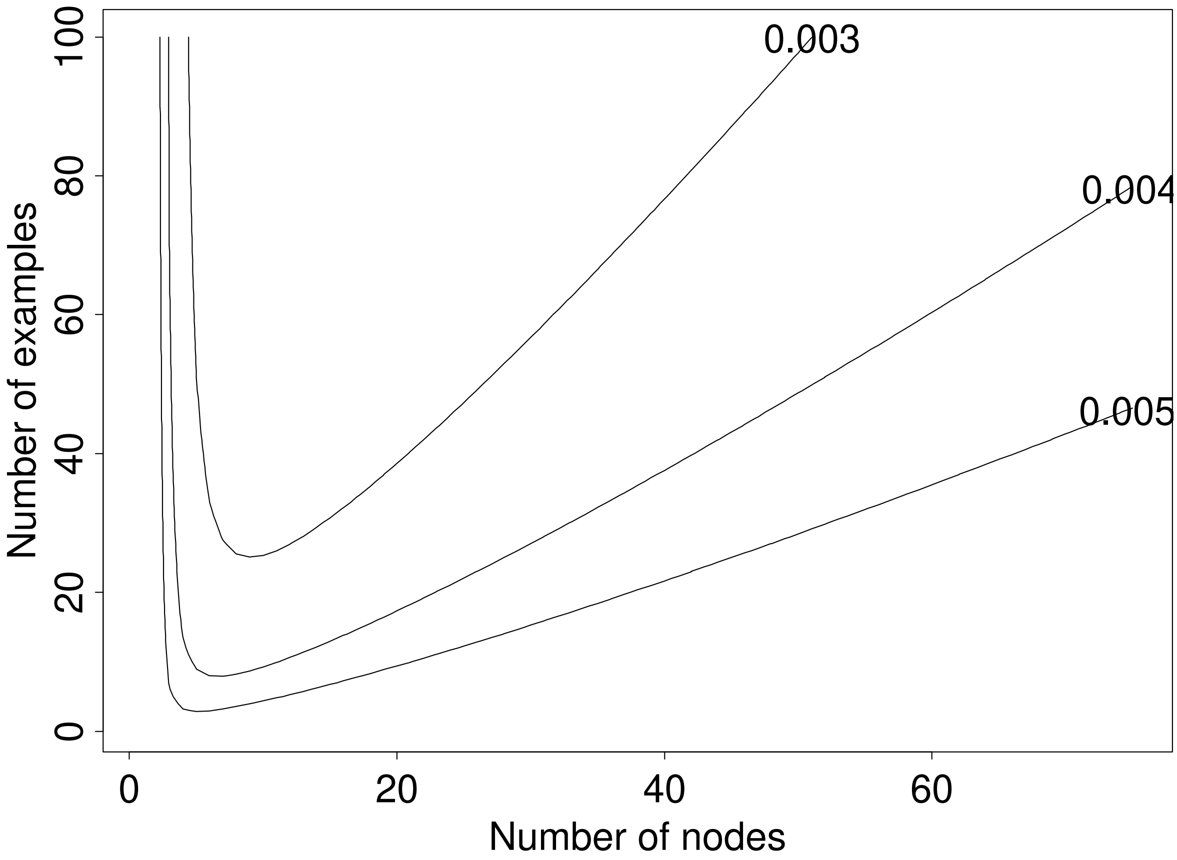

Figure 5: This figures shows various choices of (n, l) which give the same bound E (14) on the generalization error. The interesting observation is that there are an infinite number of choices for number of basis functions and number of data points all of which would guarantee the same bound on the generalization error. If (n^*, l^*) are the coordinates of the minimum of this curve, l^* is the minimum number of points necessary to achieve the error bound E with the optimal number of parameters n^*. The asymptote on the curve occurs at $n = \frac{1}{E}$, and corresponds to the case in which $l \to \infty$ and the estimation error is zero.

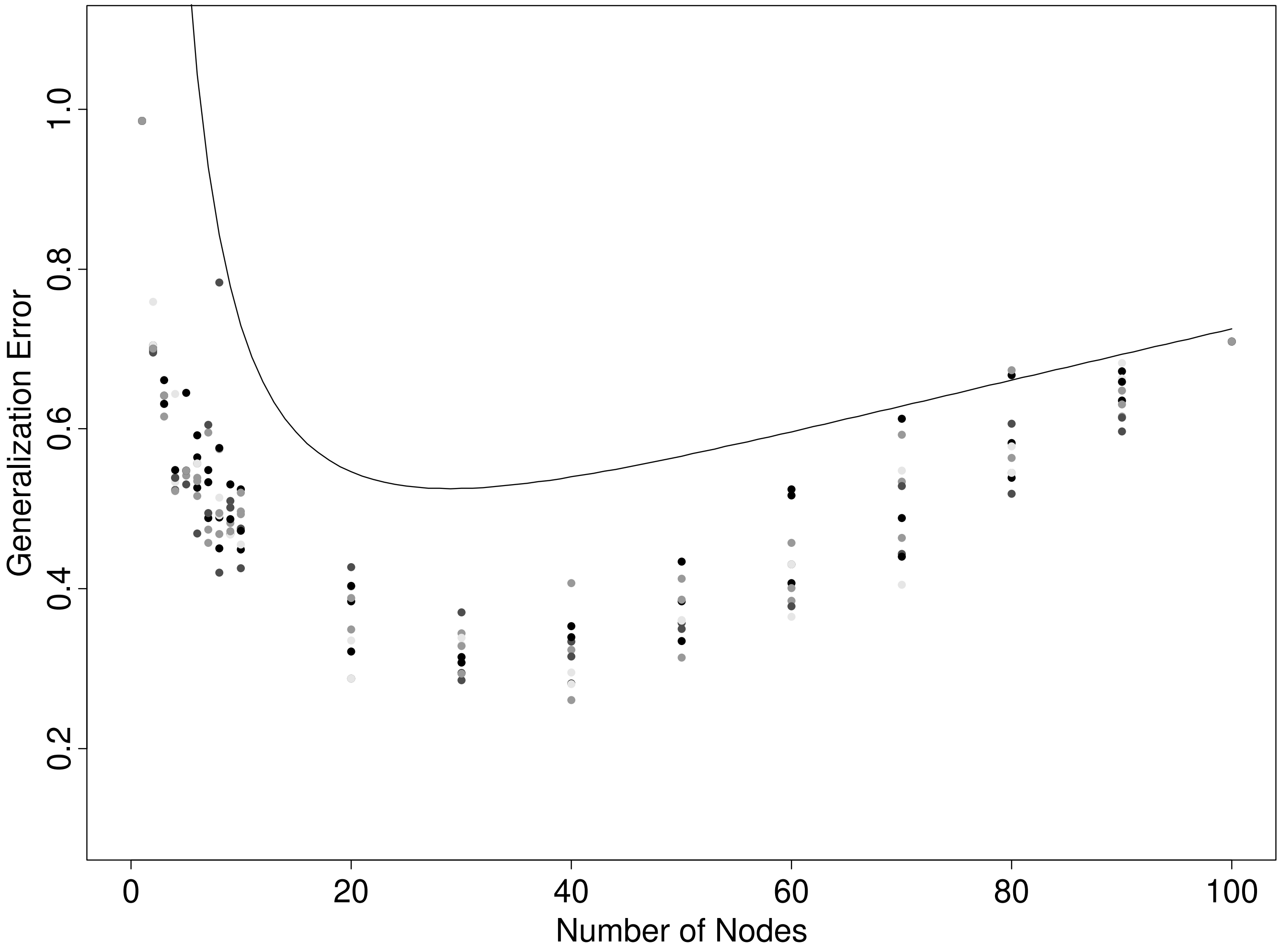

Figure 6: The generalization error is plotted as a function of the number of nodes of an RBF network (10) trained on 100 data points of a function of the type (16). For each number of parameters 10 results, corresponding to 10 different local minima, are reported. The continuous lines above the experimental data represents the bound $\frac{a}{n} + b[(nk\ln(nl) - \ln\delta)/l]^{1/2}$ of eq. (14), in which the parameters a and b have been estimated empirically, and $\delta = 10^{-6}$.

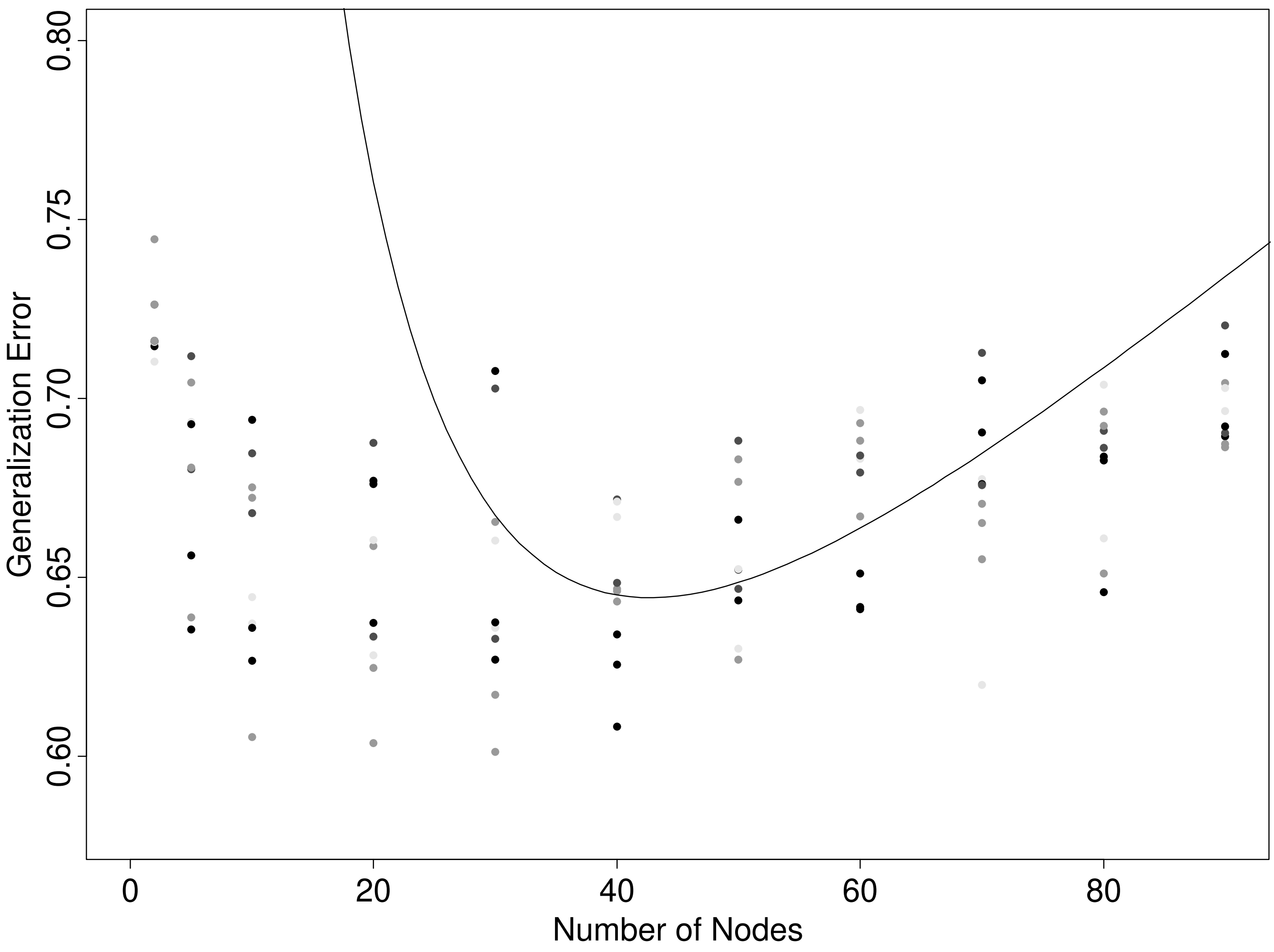

Figure 7: Everything is as in figure (6), but here the dimensionality is 6 and the number of data points is 150