1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1994	memorandum

4. TITLE AND SUBTITLE

On the Relationship between Generalization Error, Hypothesis Complexity, and Sample Complexity for Radial Basis Functions

5. FUNDING NUMBERS

N00014-92-J-1879
N00014-93-1-0385
NSF ASC19217041
N00014-93-J-0385

6. AUTHOR(S)

Partha Niyogi and Federico Girosi

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Massachusetts Institute of Technology
Artificial Intelligence Laboratory
545 Technology Square
Cambridge, Massachusetts 02139

8. PERFORMING ORGANIZATION REPORT NUMBER

AIM 1467
CBCL 88

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
Information Systems
Arlington, Virginia 22217

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

None

12a. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

In this paper, we bound the generalization error of a class of Radial Basis Function networks, for certain well defined function learningtasks, in terms of the number of parameters and number of examples. We show that the total generalization error is partly due to the insufficient representational capacity of the network (because of its finite size) and partly due to insufficient information about the target function (because of finite number of samples). We make several observations about generalization error which are valid irrespective of the approximation scheme. Our result also sheds light on ways to choose an appropriate network architecture for a particular problem.

14. SUBJECT TERMs

neural networks
sample complexity
model selection
uniform convergence
radial basis functions
non-linear regression

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED

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.

Such a scenario of course exists in a wide range of scientific disciplines. For example, in speech recognition, there might exist some functional relationship between sounds and their phonetic identities. We are given (sound, phonetic identity) pairs from which we try to infer the underlying function. This example from speech recogniton belongs to a large class of pattern classification problems where the patterns could be visual, acoustic, or tactile. In economics, it is sometimes of interest to predict the future foreign currency rates on the basis of the past time series. There might be a function which captures the dynamical relation between past and future currency rates and one typically tries to uncover this relation from data which has been appropriately processed. Similarly in medicine, one might be interested in predicting whether or not breast cancer will recur in a patient within five years after her treatment. The input space might involve dimensions like the age of the patient, whether she has been through menopause, the radiation treatment previously used etc. The output space would be single dimensional boolean taking on values depending upon whether breast cancer recurs or not. One might collect data from case histories of patients and try to uncover the underlying function.

The unknown target function is assumed to belong to some class $\mathcal{F}$ which using the terminology of computational learning theory we call 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. One can make many assumptions about how this data set is collected but a common assumption which would suffice for our purposes is 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) about the identity of the target function i.e., it comes up with a function chosen from some class, say H (the *hypothesis class*) which best fits the data and postulates this to be the target. Hypothesis classes could also be of different kinds. For example, they could be classes of boolean functions, polynomials, linear functions, spline functions and so on. One such class which is being increasingly used for learning problems is the class of feedforward networks [53],[43],[35]. A typical feedforward network is a parametrized function of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i H(\mathbf{x}; \mathbf{w}_i)$$

where $\{c_i\}_{i=1}^n$ and $\{\mathbf{w}_i\}_{i=1}^n$ are free parameters and $H(\cdot;\cdot)$ is a given, fixed function (the "activation function"). Depending on the choice of the activation function one gets different network models, such as the most common form of "neural networks", the Multilayer Perceptron [74, 18, 51, 43, 44, 30, 57, 56, 46], or the Radial Basis Functions network [14, 26, 39, 40, 58, 70, 59, 67, 66, 32, 35].

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, 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*. In this case, all the learner can hope to do is to converge to h in the limit of infinite data and so it will never recover the target. 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 a further error. This additional error (caused by finiteness of data) 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 et al [31] in the context of neural networks, and others [72, 38, 80, 75] in statistics in general.

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 total error. Second, we explore this relationship in the specific context of a particular hypothesis class. This is the class of Radial Basis function networks which can be considered to belong to the broader class of feed-forward networks. 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 [55] several researchers [18, 41, 6, 44, 15, 3, 57, 56, 46, 76] 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 assume infinite data and questions of learnability from finite data are ignored. 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 [80, 81, 82, 83] and Pollard [69], work has also been done [42, 9] 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 [42] in an excellent investigation of the sample complexity question.

In this paper, 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 White [85] and Barron [7] have provided excellent treatments of this problem for different hypothesis classes. We will mention their work at appropriate points in this paper.

The plan of the paper is as follows: in section 2 we will formalize the problem and comment on issues of a general nature. We then provide in section 3 a precise statement of a specific problem. In section 4 we present our main result, whose proof is postponed to appendix D for continuity of reading. The main result is qualified by several remarks in section 5. In section 6 we will discuss what could be the implications of our result in practice and finally we conclude in section 7 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. A summary of the most common notations and definitions used in this paper can be found in appendix A.

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, so that the independent variable will be a k-dimensional vector and the response a real number. We assume that a probability distribution $P(\mathbf{x}, y)$ is defined on $X \times Y$. P is unknown, although certain assumptions on it will be made later in this section.

The probability distribution $P(\mathbf{x}, y)$ can also be written as[1]:

$$P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x}) \ , \tag{1}$$

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 given by:

$$P(\mathbf{x}) = \int_Y dy \ P(\mathbf{x}, y) \ .$$

Expected values with respect to $P(\mathbf{x}, y)$ or $P(\mathbf{x})$ will be always indicated by $E[\cdot]$. Therefore, we will write:

$$E[g(\mathbf{x}, y)] \equiv \int_{X \times Y} d\mathbf{x}dy \ P(\mathbf{x}, y)g(\mathbf{x}, y)$$

and

$$E[h(\mathbf{x})] \equiv \int_X d\mathbf{x} \ P(\mathbf{x})h(\mathbf{x})$$

for any arbitrary function g or h.

2.2 Learning from Examples and Estimators

The framework described above can be used to model the fact that in the real world we often have to deal with sets of variables that are related by a probabilistic relationship. For example, y could be the measured torque at a particular joint of a robot arm, and $\mathbf{x}$ the set of angular position, velocity and acceleration of the joints of the arm in a particular configuration. The relationship between $\mathbf{x}$ and y is probabilistic because there is noise affecting the measurement process, so that two different torques could be measured given the same configuration.

In many cases we are provided with *examples* of this probabilistic relationship, that is with a data set D_l, obtained by sampling l times the set $X \times Y$ according to $P(\mathbf{x}, y)$:

$$D_l \equiv \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^{l} \ .$$

From eq. (1) 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 $P(y|\mathbf{x})$. In the robot arm example described above, it would mean that one could move the robot arm into

[1] Note that we are assuming that the conditional distribution exists, but this is not a very restrictive assumption.

a random configuration $\mathbf{x}_1$, measure the corresponding torque y_1, and iterate this process l times.

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. For example, given a certain configuration of the robot arm, we would like to estimate the corresponding torque.

Formally, we define an *estimator* to be any function $f : X \rightarrow Y$. 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. Thus we have a probabilistic relation between $\mathbf{x}$ and y. One can think of this as an underlying deterministic relation corrupted with noise. Hopefully a good estimator will be able to recover this relation.

2.3 The Expected Risk and the Regression Function

In the previous section we explained the problem of learning from examples and stated that this is the same as the problem of finding the best estimator. To make sense of this statement, 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 measure[2] of the error of the estimator f at the point $\mathbf{x}$ is:

$$(y - f(\mathbf{x}))^2 .$$

In the example of the robot arm, $f(\mathbf{x})$ is our estimate of the torque corresponding to the configuration $\mathbf{x}$, and y is the measured torque of that configuration. 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.

Given this particular measure as our yardstick to evaluate different estimators, we are now interested in finding the estimator that minimizes the expected risk. In order to proceed we need to specify its domain of definition $\mathcal{F}$. Then using the expected risk as a criterion, we could obtain the best element of $\mathcal{F}$. Depending on the properties of the unknown probability distribution $P(\mathbf{x}, y)$ one could make different choices for $\mathcal{F}$. We will assume in the following that $\mathcal{F}$ is some space of differentiable functions. For example, $\mathcal{F}$ could be a space of functions with a certain number of bounded derivatives (the spaces $\Lambda^m(R^d)$ defined in appendix A), or a Sobolev space of functions with a certain number of derivatives in L_p (the spaces $H^{m,p}(R^d)$ defined in appendix A).

[2]Note that this is the familiar squared-error and when averaged over its domain yields the mean squared error for a particular estimator, a very common choice. However, it is useful to remember that there could be other choices as well.

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 B):

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

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

$$f_0(\mathbf{x}) \equiv \int_Y dy \; yP(y|\mathbf{x}) . \quad (3)$$

From eq. (2) 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,

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

However, it is also clear that even the regression function will make an error equal to $E[(y - f_0(\mathbf{x}))^2]$, that is the variance of the response given a certain value for the independent variable, averaged over the values the independent variable can take. While the first term in eq. (2) depends on the choice of the estimator f, the second term is an intrinsic limitation that comes from the fact that the independent variable $\mathbf{x}$ does not determine uniquely the response y.

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. Thus we have some large class of functions $\mathcal{F}$ to which the target function f_0 belongs. We obtain noisy data of the form $(\mathbf{x}, y)$ where $\mathbf{x}$ has the distribution $P(\mathbf{x})$ and for each $\mathbf{x}$, y is a random variable with mean $f_0(\mathbf{x})$ and distribution $P(y|\mathbf{x})$. We note that y can be viewed as a deterministic function of $\mathbf{x}$ corrupted by noise. If one assumes the noise is additive, we can write $y = f_0(\mathbf{x}) + \eta_x$ where η_x[3] is zero-mean with distribution $P(y|\mathbf{x})$. We choose an estimator on the basis of the data set and we hope that it is close to the regression (target) function. 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 [36, 45, 71].

2.4 The Empirical Risk

If the expected risk functional $I[f]$ were known, one could compute the regression function by simply finding its minimum in $\mathcal{F}$, that would make the whole learning problem considerably easier. What makes the problem difficult and interesting is that in practice $I[f]$ is unknown because $P(\mathbf{x}, y)$ is unknown. Our only source of information is the data set D_l which consists of l independent random samples of $X \times Y$ drawn according to $P(\mathbf{x}, y)$. Using this data set, the expected risk can be approximated by the *empirical risk* I_{emp}:

[3]Note that the standard regression problem often assumes η_x is independent of x. Our case is distribution free because we make no assumptions about the nature of η_x.

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

For each given estimator f, the empirical risk is a random variable, and under fairly general assumptions[4], by the law of large numbers [23] it converges in probability to the expected risk as the number of data points goes to infinity:

$$\lim_{l\to\infty} P\{|I[f] - I_{\text{emp}}[f]| > \varepsilon\} = 0 \quad \forall \varepsilon > 0 . \qquad (4)$$

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. For the error metric we have used, this yields the least-squares error estimator. However, eq. (4) 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 [81, 82, 83] the notion of *uniform convergence* in probability has to be introduced, and it will be discussed in other parts of this paper.

2.5 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

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

However this problem is clearly 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 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, multilayer perceptrons with n threshold units and so on. Therefore, we choose as approximation to the regression function the function $\hat{f}_{n,l}$ defined as:[5]

$$\hat{f}_{n,l} \equiv \arg\min_{f\in H_n} I_{\text{emp}}[f] . \qquad (5)$$

Thus, for example, if H_n is the class of functions which can be represented as $f = \sum_{\alpha=1}^{n} c_\alpha H(\mathbf{x};\mathbf{w}_\alpha)$ then eq. (5) can be written as

$$\hat{f}_{n,l} \equiv \arg\min_{c_\alpha,\mathbf{w}_\alpha} I_{\text{emp}}[f] .$$

A number of observations need to be made here. First, if the class $\mathcal{F}$ is small (typically in the sense of bounded VC-dimension or bounded metric entropy [69]), then the problem is not necessarily ill-posed and we do not have to go through the process of using the sets H_n. However, as has been mentioned already, for most interesting choices of $\mathcal{F}$ (e.g. classes of functions in Sobolev spaces, continuous functions etc.) the problem might be ill posed. However, this might not be the only reason for using the classes H_n. It might be the case that that is all we have or for some reason it is something we would like to use. For example, one might want to use a particular class of feed-forward networks because of ease of implementation in VLSI. Also, if we were to solve the function learning problem on a computer as is typically done in practice, then the functions in $\mathcal{F}$ have to be represented somehow. We might consequently use H_n as a representation scheme. It should be pointed out that the sets H_n and $\mathcal{F}$ have to be matched with each other. For example, we would hardly use polynomials as an approximation scheme when the class $\mathcal{F}$ consists of indicator functions or for that matter use threshold units when the class $\mathcal{F}$ contains continuous functions. In particular, if we are to recover the regression function, H must be dense in $\mathcal{F}$. 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, if we were to use multilayer perceptrons, this line of questioning would lead us to identify the class of problems which can be effectively solved by them.

Thus, we see that in principle we would like to minimize $I[f]$ over the large class $\mathcal{F}$ obtaining thereby the

[4] For example, assuming the data is independently drawn and $I[f]$ is finite.

[5] Notice that we are implicitly assuming that the problem of minising $I_{\text{emp}}[f]$ over H_n has a solution, which might not be the case. However the quantity

$$E_{n,l} \equiv \inf_{f\in H_n} I_{\text{emp}}[f]$$

is always well defined, and we can always find a function $\hat{f}_{n,l}$ for which $I_{\text{emp}}[\hat{f}_{n,l}]$ is arbitrarily close to $E_{n,l}$. It will turn out that this is sufficient for our purposes, and therefore we will continue, assuming that $\hat{f}_{n,l}$ is well defined by eq. (5)

regression function f_0. What we do in practice is to minimize the empirical risk $I_{\text{emp}}[f]$ 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.

2.6 Bounding the Generalization Error

At this stage it might be worthwhile to review and remark on some general features of the problem of learning from examples. Let us remember that our goal is to minimize the expected risk $I[f]$ over the set $\mathcal{F}$. If we were to use a finite number of parameters, then we have already seen that *the best we could possibly do is to minimize* our functional over the set H_n, yielding the estimator f_n:

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

However, not only is the parametrization limited, but the data is also finite, and we can only minimize the empirical risk I_{emp}, obtaining as our final estimate the function $\hat{f}_{n,l}$. Our goal is to bound the distance from $\hat{f}_{n,l}$ that is our solution, from f_0, that is the "optimal" solution. If we choose to measure the distance in the $L^2(P)$ metric (see appendix A), the quantity that we need to bound, that we will call *generalization error*, is:

$$E[(f_0 - \hat{f}_{n,l})^2] = \int_X dx \; P(x)(f_0(x) - \hat{f}_{n,l}(x))^2 =$$

$$= \|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)}$$

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 cause of error comes from the fact that we are trying to approximate an infinite dimensional object, the regression function $f_0 \in \mathcal{F}$, with a finite number of parameters. We call this error *the approximation error*, and we measure it by the quantity $E[(f_0 - f_n)^2]$, that is the $L_2(P)$ distance between the best function in H_n and the regression function. The approximation error can be expressed in terms of the expected risk using the decomposition (2) as

$$E[(f_0 - f_n)^2] = I[f_n] - I[f_0] \, . \qquad (6)$$

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. The natural framework to study it is approximation theory, that abound with bounds on the approximation error for a variety of choices of H_n and $\mathcal{F}$. 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 shown in figure (1), 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 [18, 44, 6, 8, 30, 57, 56] 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* [80]. 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 [81, 82, 80, 83] Pollard [69], Dudley [24], and Haussler [42]. Under a variety of different hypothesis it is possible to prove that, with probability $1 - \delta$, a bound of this form is valid:

$$|I_{\text{emp}}[f] - I[f]| \leq \omega(l, n, \delta) \quad \forall f \in H_n \qquad (7)$$

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, the generalization error can be bounded by the sum of the two:

Statement 2.1 *The following inequality holds:*

$$\|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)} \leq \varepsilon(n) + 2\omega(l, n, \delta) \ . \qquad (8)$$

Proof: using the decomposition of the expected risk (2), the generalization error can be written as:

$$\|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)} = E[(f_0 - \hat{f}_{n,l})^2] = I[\hat{f}_{n,l}] - I[f_0] \ . \ (9)$$

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

$$E[(f_0 - \hat{f}_{n,l})^2] \leq |I[f_n] - I[f_0]| + |I[f_n] - I[\hat{f}_{n,l}]| \ . \ (10)$$

In the first term of the right hand side of the previous inequality we recognize the approximation error (6). If a bound of the form (7) is known for the generalization error, it is simple to show (see appendix (C) 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 □.

Thus we see that the generalization error has two components: one, bounded by $\varepsilon(n)$, is related to the approximation power of the class of functions $\{H_n\}$, and is studied in the framework of approximation theory. The second, bounded by $\omega(l, n, \delta)$, is related to the difficulty of estimating the parameters given finite data, and is studied in the framework of statistics. Consequently, results from both these fields are needed in order to provide an understanding of the problem of learning from examples. Figure (1) also shows a picture of the problem.

2.7 A Note on Models and Model Complexity

From the form of eq. (8) the reader will quickly 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 [29, 75, 52, 73, 4, 28, 17] 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 might be true even when considering qualitatively different models and as an illustrative example let us consider two kinds of models we might use to learn regression functions in the space of bounded continuous functions. The class of linear models, i.e., the class of functions which can be expressed as $f = \mathbf{w} \cdot \mathbf{x} + \theta$, do not have much approximating power and consequently their approximation error is rather high. However, their estimation error is quite low. The class of models which can be expressed in the form $H = \sum_{i=1}^{n} c_i \sin(\mathbf{w}_i \cdot \mathbf{x} + \theta_i)$ have higher approximating power [47] resulting in low approximation errors. However this class has an infinite VC-dimension [82] and its estimation error can not therefore be bounded.

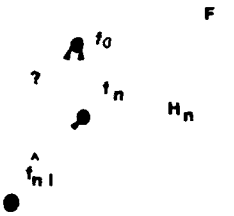

Figure 1: This figure shows a picture of the problem. The outermost circle represents the set F. Embedded in this are the nested subsets, the H_n's. f_0 is an arbitrary target function in $\mathcal{F}$, f_n is the closest element of H_n and $\hat{f}_{n,l}$ is the element of H_n which the learner hypothesizes on the basis of data.

So far we have provided a very general characterization of this problem, without stating what the sets $\mathcal{F}$ and H_n are. As we have already mentioned before, the set $\mathcal{F}$ could be a set of bounded differentiable or integrable functions, and H_n could be polynomials of degree n, spline functions with n knots, multilayer perceptrons with n hidden units or any other parametric approximation scheme with n parameters. 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. (8).

3 Stating the Problem for Radial Basis Functions

As mentioned before the problem of learning from examples reduces to estimating some target function from a set X to a set Y. In most practical cases, such as character recognition, motor control, time series prediction, the set X is the k-dimensional Euclidean space R^k, and the set Y is some subset of the real line, that for our purposes we will assume to be the interval $[-M, M]$, where M is some positive number. In fact, there is a probability distribution $P(\mathbf{x}, y)$ defined on the space $R^k \times [-M, M]$ according to which the labelled examples are drawn independently at random, and from which we try to estimate the regression (target) function. It is clear that the regression function is a real function of k variables.

In this paper we focus our attention on the Radial Ba-

sis Functions approximation scheme (also called Hyper-Basis Functions [67]). This is the class of approximating functions that can be written as:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \beta_i G(\mathbf{x} - \mathbf{t}_i)$$

where G is some given basis function and the β_i and the $\mathbf{t}_i$ 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. We will see later that a favorable approximation bound can be obtained if we assume that the class of functions $\mathcal{F}$ to which the regression function belongs is defined as follows:

$$\mathcal{F} \equiv \{f \in L_2(R^k) | f = \lambda * G, |\lambda|_{R^k} \leq M\} . \qquad (11)$$

Here λ is a signed Radon measure on the Borel sets of R^k, G is a gaussian function with range in $[0, V]$, the symbol $*$ stands for the convolution operation, $|\lambda|_{R^k}$ is the total variation[6] of the measure λ and M is a positive real number. We point out that the class $\mathcal{F}$ is non-trivial to learn in the sense that it has infinite pseudo-dimension [69].

In order to obtain an estimation bound we need the approximating class to have bounded variation, and the following constraint will be imposed:

$$\sum_{i=1}^{n} |\beta_i| \leq M .$$

We will see in the proof that 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

$$f(\mathbf{x}) = \sum_{i=1}^{n} \beta_i G(\mathbf{x} - \mathbf{t}_i), \ \sum_{i=1}^{n} |\beta_i| \leq M \ , \ \ \mathbf{t}_i \in R^k \quad (12)$$

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:

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

Notice that assumption that the regression function

$$f_0(\mathbf{x}) \equiv E[y|\mathbf{x}]$$

belongs to the class $\mathcal{F}$ correspondingly implies an assumption on the probability distribution $P(y|\mathbf{x})$, viz., that P must be such that $E[y|\mathbf{x}]$ belongs to $\mathcal{F}$. Notice also that since we assumed that Y is a closed interval, we are implicitly assuming that $P(y|\mathbf{x})$ has compact support.

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.

4 Main Result

The main theorem is:

Theorem 4.1 *For any $0 < \delta < 1$, for n nodes, l data points, input dimensionality of k, and $H_n, \mathcal{F}, f_0, \hat{f}_{n,l}$ also as defined in the statement of the problem above, with probability greater than $1 - \delta$,*

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

Proof: The proof requires us to go through a series of propositions and lemmas which have been relegated to appendix (D) for continuity of ideas.□

5 Remarks

There are a number of comments we would like to make on the formulation of our problem and the result we have obtained. There is a vast body of literature on approximation theory and the theory of empirical risk minimization. In recent times, some of the results in these areas have been applied by the computer science and neural network community to study formal learning models. Here we would like to make certain observations about our result, suggest extensions and future work, and to make connections with other work done in related areas.

5.1 Observations on the Main Result

- The theorem has a PAC[79] like setting. It tells us that if we draw enough data points (labelled examples) and have enough nodes in our Radial Basis Functions network, we can drive our error arbitrarily close to zero with arbitrarily high probability. Note however that our result is not entirely distribution-free. Although no assumptions are made on the form of the underlying distribution, we do have certain constraints on the kinds of distributions for which this result holds. In particular, the distribution is such that its conditional mean $E[y|\mathbf{x}]$ (this is also the regression function $f_0(\boldsymbol{x})$) must belong to a the class of functions $\mathcal{F}$ defined by eq. (11). Further the distribution $P(y|\mathbf{x})$ must have compact support [7].

[6] A signed measure λ can be decomposed by the Hahn-Jordan decomposition into $\lambda = \lambda^+ - \lambda^-$. Then $|\lambda| = \lambda^+ + \lambda^-$ is called the total variation of λ. See Dudley [23] for more information.

[7] This condition, that is related to the problem of large deviations [80], could be relaxed, and will be subject of further investigations.

- The error bound consists of two parts, one ($O(1/n)$) coming from approximation theory, and the other $O(((nk\ln(nl) + \ln(1/\delta))/l)^{1/2})$ from statistics. It is noteworthy that for a given approximation scheme (corresponding to $\{H_n\}$), a certain class of functions (corresponding to $\mathcal{F}$) suggests itself. So we have gone from the class of networks to the class of problems they can perform as opposed to the other way around, i.e., from a class of problems to an optimal class of networks.
- This sort of a result implies that if we have the prior knowledge that f_0 belongs to class $\mathcal{F}$, then by choosing the number of data points, l, and the number of basis functions, n, appropriately, we can drive the misclassification error arbitrarily close to Bayes rate. In fact, for a fixed amount of data, even before we have started looking at the data, we can pick a starting architecture, i.e., the number of nodes, n, for optimal performance. After looking at the data, we might be able to do some structural risk minimization [80] to further improve architecture selection. For a fixed architecture, this result sheds light on how much data is required for a certain error performance. Moreover, it allows us to choose the number of data points and number of nodes simultaneously for guaranteed error performances. Section 6 explores this question in greater detail.

5.2 Extensions

- There are certain natural extensions to this work. We have essentially proved the consistency of the estimated network function $\hat{f}_{n,l}$. In particular we have shown that $\hat{f}_{n,l}$ converges to f_0 with probability 1 as l and n grow to infinity. It is also possible to derive conditions for almost sure convergence. Further, we have looked at a specific class of networks ($\{H_n\}$) which consist of weighted sums of Gaussian basis functions with moving centers but fixed variance. This kind of an approximation scheme suggests a class of functions $\mathcal{F}$ which can be approximated with guaranteed rates of convergence as mentioned earlier. We could prove similar theorems for other kinds of basis functions which would have stronger approximation properties than the class of functions considered here. The general principle on which the proof is based can hopefully be extended to a variety of approximation schemes.
- We have used notions of metric entropy and covering number [69, 24] in obtaining our uniform convergence results. Haussler [42] uses the results of Pollard and Dudley to obtain uniform convergence results and our techniques closely follow his approach. It should be noted here that Vapnik [80] deals with exactly the same question and uses the VC-dimension instead. It would be interesting to compute the VC-dimension of the class of networks and use it to obtain our results.
- 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.
- 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 favours less complex (according to the defined notion of complexity) functions over more complex ones. For example the regularization approach [77, 68, 84] 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. It would be interesting to obtain convergence conditions and rates for such schemes. Choice of an optimal λ is an interesting question in regularization techniques and typically cross-validation or other heuristic schemes are used. A result on convergence rate potentially offers a principled way to choose λ.
- Structural risk minimization 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[8]. 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 [37]. 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 [2, 62] 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. It is conceivable that such a scheme would grow to minimal parametrization (or closer to it at any rate) and require less data than classical structural risk minimization.
- 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

[8] Neither does regularization for that matter. The question of minimal parametrization is related to that of order determination of systems, a very difficult problem!

computational complexity of obtaining even an approximate solution to the minimization problem is an interesting one and results from computer science [49, 12] suggest that it might in general be NP-hard.

5.3 Connections with Other Results

- 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 [10, 42, 61] in characterizing the sample complexity of finite sized networks. Of these, it is worthwhile to mention again the work of Haussler [42] 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. Later we will characterize the trade-off between hypothesis complexity and error rate. Others [27, 63] attempt to characterize the generalization abilities of feed-forward networks using theoretical formalizations from statistical mechanics. Yet others [13, 60, 16, 1] attempt to obtain empirical bounds on generalization abilities.
- This is an attempt to obtain rate-of-convergence bounds in the spirit of Barron's work [5], 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. It should be mentioned here that his proof relies on a discretization of the networks into a countable family, while no such assumption is made here.
- It would be worthwhile to make a reference to Geman's paper [31] which talks of the Bias-Variance dilemma. This is another way of formulating the trade-off between the approximation error and the estimation error. As the number of parameters (proportional to n) increases, the bias (which can be thought of as analogous to the approximation error) of the estimator decreases and its variance (which can be thought of as analogous to the estimation error) increases for a fixed size of the data set. Finding the right bias-variance trade-off is very similar in spirit to finding the trade-off between network complexity and data complexity.
- Given the class of radial basis functions we are using, a natural comparison arises with kernel regression [50, 22] 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.
- 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 [71, 45, 36] 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.

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

We have stated our main result in a particular form. We have provided a provable upper bound on the error (in the $\|\cdot\|_{L^2(P)}$ metric) in terms of the number of examples and the number of basis functions used. Further we have provided the order of the convergence and have not stated the constants involved. The same result could be stated in other forms and has certain implications. It provides us rates at which the number of basis functions (n) should increase as a function of the number of examples (l) in order to guarantee convergence(Section 6.1). It also provides us with the trade-offs between the two as explored in Section 6.2.

6.1 Rate of Growth of n for Guaranteed Convergence

From our theorem (4.1) 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

$$\lim_{l\to+\infty} \omega(l, n, \delta) =$$

$$= \lim_{l\to+\infty} O\left(\left[\frac{l^r k \ln(l^{r+1}) + \ln(1/\delta)}{l}\right]^{1/2}\right) =$$

$$= \lim_{l\to+\infty} l^{r-1}\ln l\ .$$

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

6.2 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. This phenomenon is evident from figure (2), where we plotted the generalization error 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.

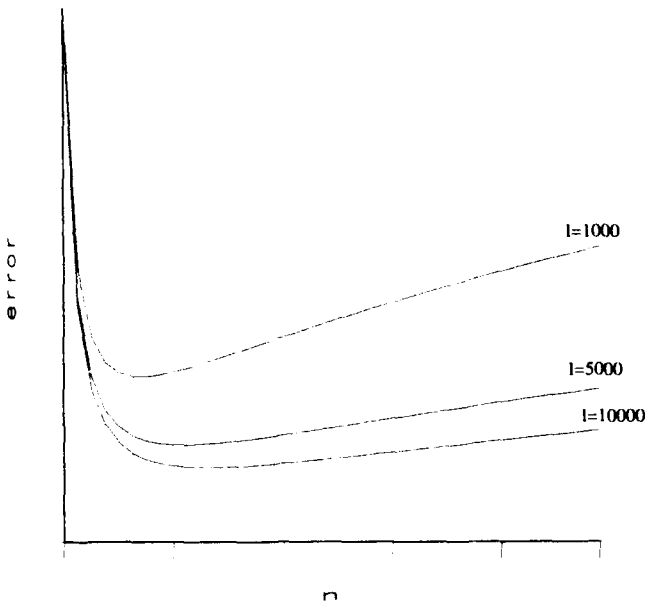

Figure 2: Bound on the generalization error as a function of the number of basis functions n keeping the sample size l fixed. This has been plotted for a few different choices of sample size. Notice how the generalization error goes through a minimum for a certain value of n. This would be an appropriate choice for the given (constant) data complexity. Note also that the minimum is broader for larger l, that is, an accurate choice of n is less critical when plenty of data is available.

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 (4.1) in the previous equation, and setting all the constants to 1 for simplicity, we obtain:

$$\frac{\partial}{\partial n}\left[\frac{1}{n} + \left(\frac{nk\ln(nl) - \ln(\delta)}{l}\right)^{\frac{1}{2}}\right] = 0 .$$

Performing the derivative the expression above can be written as

$$\frac{1}{n^2} = \frac{1}{2}\left[\frac{kn\ln(nl) - \ln\delta}{l}\right]^{-\frac{1}{2}} \frac{k}{l}[\ln(nl) + 1] .$$

We now make the assumption that l is big enough to let us perform the approximation $\ln(nl) + 1 \approx \ln(nl)$. Moreover, we assume that

$$\frac{1}{\delta} << (nl)^{nk}$$

in such a way that the term including δ in the equation above is negligible. After some algebra we therefore conclude that the optimal number of parameters $n^*(l)$ satisfies, for large l, the equation:

$$n^*(l) = \left[\frac{4l}{k\ln(n^*(l)l)}\right]^{\frac{1}{3}} .$$

From this equation is clear that n^* is roughly proportional to a power of l, and therefore we can neglect the factor n^* in the denominator of the previous equation, since it will only affect the result by a multiplicative constant. Therefore we conclude that the optimal number of parameters $n^*(l)$ for a given number of examples behaves as

$$n^*(l) \propto \left[\frac{l}{k\ln l}\right]^{\frac{1}{3}} . \qquad (14)$$

In order to show that this is indeed the optimal rate of growth we reported in figure (3) the generalization error as function of the number of examples l for different rate of growth of n, that is setting $n = l^r$ for different values of r. Notice that the exponent $r = \frac{1}{3}$, that is very similar to the optimal rate of eq. (14), performs better than larger ($r = \frac{1}{2}$) and smaller ($r = \frac{1}{10}$) exponents.

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 rate (ϵ), there are various choices of n and l which are satisfactory. Fig. 4 shows n as a function of l, in other words (l, n) pairs which yield the same error rate with the same confidence.

If data are expensive for us, we could operate in region A of the curve. If network size is expensive we could operate in region B of the curve. In particular the economics of trading off network and data complexity would yield a suitable point on this curve and thus would allow us to choose the right combination of n and l to solve our regression problem with the required accuracy and confidence.

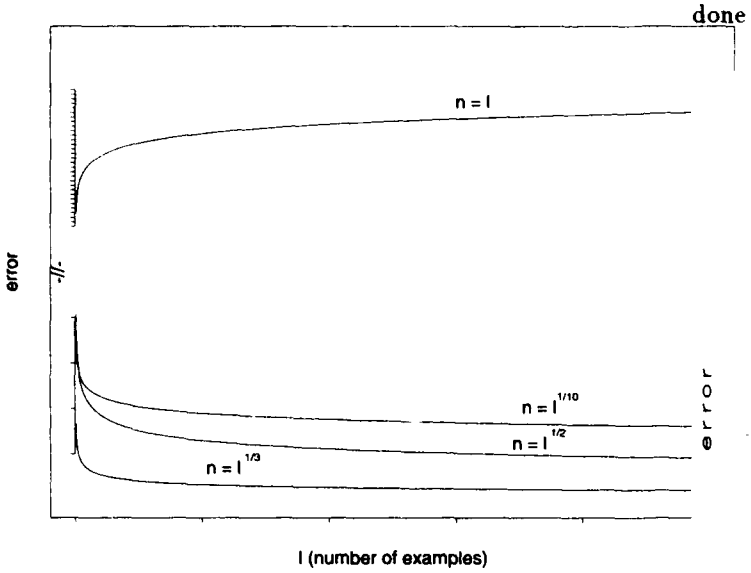

Figure 3: The bound on the generalization error as a function of the number of examples for different choices of the rate at which network size n increases with sample size l. Notice that if $n = l$, then the estimator is not guaranteed to converge, i.e., the bound on the generalization error diverges. While this is a distribution free-upper bound, we need distribution-free lower bounds as well to make the stronger claim that $n = l$ will never converge.

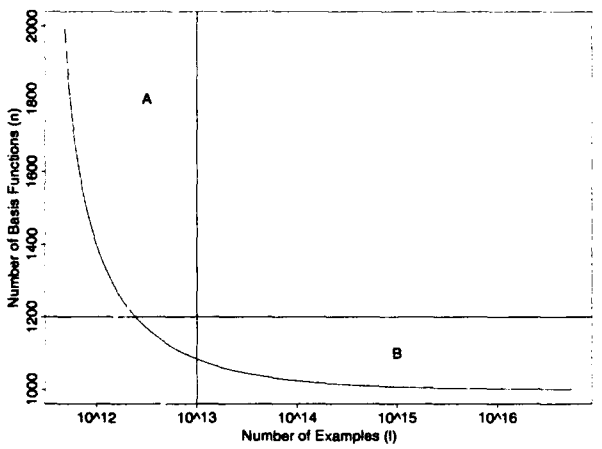

Figure 4: This figures shows various choices of (l, n) which give the same generalization error. The x-axis has been plotted on a log scale. The interesting observation is that there are an infinite number of choices for number of basis functions and number of data points all of which would guarantee the same generalization error (in terms of its worst case bound).

Of course we could also plot the error as a function of data size l for a fixed network size (n) and this has been done for various choices of n in Fig. 5.

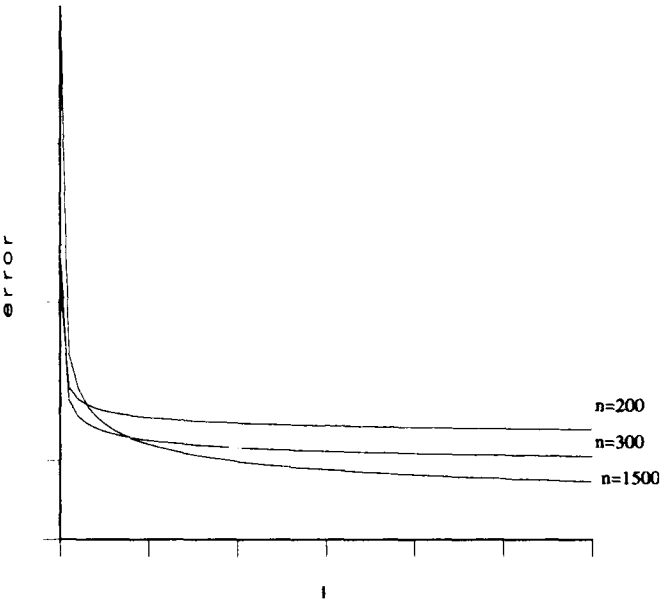

Figure 5: The generalization error as a function of number of examples keeping the number of basis functions (n) fixed. This has been done for several choices of n. As the number of examples increases to infinity the generalization error asymptotes to a minimum which is not the Bayes error rate because of finite hypothesis complexity (finite n).

We see as expected that the error monotonically decreases as a function of l. However it asymptotically decreases not to the Bayes error rate but to some value above it (the approximation error) which depends upon the the network complexity.

Finally figure (6) shows the result of theorem (4.1) in a 3-dimensional plot. The generalization error, the network size, and the sample size are all plotted as a function of each other.

7 Conclusion

For the task of learning some unknown function from labelled 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

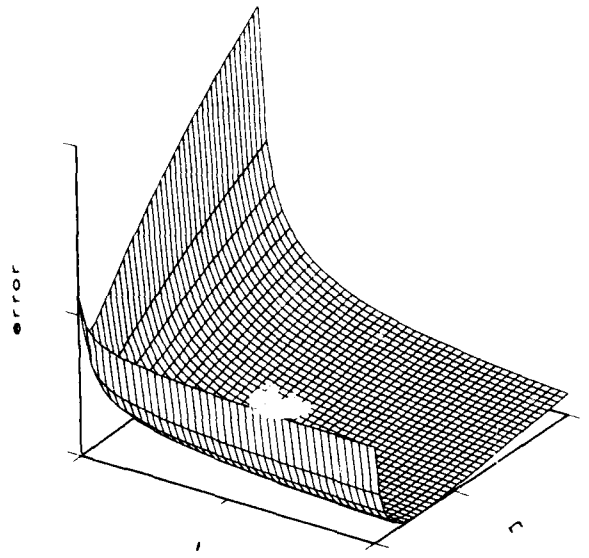

Figure 6: The generalization error, the number of examples (l) and the number of basis functions (n) as a function of each other.

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 longstanding interest in statistics and we provide connections between our work and other work in the field.

Acknowledgments We are grateful to T. Poggio and B. Caprile for useful discussions and suggestions.

A Notations

- $\mathcal{A}$: a set of functions defined on S such that, for any $a \in \mathcal{A}$,

$$0 \leq a(\xi) \leq U^2 \quad \forall \xi \in S \ .$$

- $\mathcal{A}_{\tilde{\xi}}$: the restriction of $\mathcal{A}$ to the data set, see eq. (22).
- $\mathcal{B}$: it will usually indicate the set of all possible l-dimensional Boolean vectors.
- B: a generic ϵ-separated set in S.
- $\mathcal{C}(\epsilon, \mathcal{A}, d_{L^1})$: the metric capacity of a set $\mathcal{A}$ endowed with the metric $d_{L^1(P)}$.
- $d(\cdot,\cdot)$: a metric on a generic metric space S.
- $d_{L^1}(\cdot,\cdot)$, $d_{L^1(P)}(\cdot,\cdot)$: L^1 metrics in vector spaces. The definition depends on the space on which the metric is defined (k-th dimensional vectors, real valued functions, vector valued functions).
  1. In a vector space R^k we have

$$d_{L^1}(\mathbf{x},\mathbf{y}) = \frac{1}{l}\sum_{\mu=1}^{l} |x^\mu - y^\mu|$$

  where $\mathbf{x}$, $\mathbf{y} \in R^k$, x^μ and y^μ denote their μ-th components.
  2. In an infinite dimensional space $\mathcal{F}$ of real valued functions in k variables we have

$$d_{L^1(P)}(f,g) = \int_{R^k} |f(\mathbf{x}) - g(\mathbf{x})| dP(\mathbf{x})$$

  where $f, g \in \mathcal{F}$ and $dP(\mathbf{x})$ is a probability measure on R^k.
  3. In an infinite dimensional space $\mathcal{F}$ of functions in k variables with values in R^n we have

$$d_{L^1(P)}(\mathbf{f},\mathbf{g}) = \frac{1}{n}\sum_{i=1}^{n} \int_{R^k} |f_i(\mathbf{x}) - g_i(\mathbf{x})| dP(\mathbf{x})$$

  where
  $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots f_i(\mathbf{x}), \ldots f_n(\mathbf{x}))$, $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots g_i(\mathbf{x}), \ldots g_n(\mathbf{x}))$ are elements of $\mathcal{F}$ and $dP(\mathbf{x})$ is a probability measure on R^k.
- D_l: it will always indicate a data set of l points:

$$D_l \equiv \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^{l} \ .$$

  The points are drawn according to the probability distribution $P(\mathbf{x}, y)$.
- $E[\cdot]$: it denotes the expected value with respect to the probability distribution $P(\mathbf{x}, y)$. For example

$$I[f] = E[(y - f(\mathbf{x}))^2] \ ,$$

and

$$\|f_0 - f\|^2_{L^2(P)} = E[(f_0(\mathbf{x}) - f(\mathbf{x}))^2] \ .$$

- f: a generic estimator, that is any function from X to Y:

$$f : X \Rightarrow Y \ .$$

- $f_0(\mathbf{x})$: the regression function, it is the conditional mean of the response given the predictor:

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

It can also be defined as the function that minimizes the expected risk $I[f]$ in $\mathcal{U}$, that is

$$f_0(\mathbf{x}) \equiv \arg \inf_{f \in \mathcal{U}} I[f] \ .$$

Whenever the response is obtained sampling a function h in presence of zero mean noise the regression functior coincides with the sampled function h.

- f_n: it is the function that minimizes the expected risk $I[f]$ in H_n:

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

Since

$$I[f] = \|f_0 - f\|^2_{L^2(P)} + I[f_0]$$

f_n it is also the best $L^2(P)$ approximation to the regression function in H_n (see figure 1).

- $\hat{f}_{n,l}$: is the function that minimizes the empirical risk $I_{\text{emp}}[f]$ in H_n:

$$\hat{f}_{n,l} \equiv \arg \inf_{f \in H_n} I_{\text{emp}}[f]$$

In the neural network language it is the output of the network after training has occurred.

- $\mathcal{F}$: the space of functions to which the regression function belongs, that is the space of functions we want to approximate.

$$\mathcal{F} : X \Rightarrow Y$$

where $X \in R^d$ and $Y \in R$. $\mathcal{F}$ could be for example a set of differentiable functions, or some Sobolev space $H^{m,p}(R^k)$

- $\mathcal{G}$: it is a class of functions of k variables

$$g : R^k \rightarrow [0, V]$$

defined as

$$\mathcal{G} == \{g : g(\mathbf{x}) = G(\|\mathbf{x} - \mathbf{t}\|), \ \mathbf{t} \in R^k\}.$$

where G is the gaussian function.

- G_1: it is a $k+2$-dimensional vector space of functions from R^k to R defined as

$$G_1 \equiv \text{span}\{1, x^1, x^2, \cdot, x^k, \|\mathbf{x}\|^2\}$$

where $\mathbf{x} \in R^k$ and x^μ is the μ-th component of the vector $\mathbf{x}$.

- G_2: it is a set of real valued functions in k variables defined as

$$G_2 = \{\alpha e^{-f} : f \in G_1, \ \ \alpha = \frac{1}{\sqrt{2\pi}\sigma}\}$$

where σ is the standard deviation of the Gaussian G.

- H_I: it is a class of vector valued functions

$$\mathbf{g}(\mathbf{x}) : R^k \rightarrow R^n$$

of the form

$$\mathbf{g}(\mathbf{x}) = (G(\|\mathbf{x} - \mathbf{t}_1\|), G(\|\mathbf{x} - \mathbf{t}_2\|), \ldots, G(\|\mathbf{x} - \mathbf{t}_n\|))$$

where G is the gaussian function and the $\mathbf{t}_i$ are arbitrary k-dimensional vectors.

- H_F: it is a class of real valued functions in n variables:

$$f : [0, V]^n \rightarrow R$$

of the form

$$f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x}$$

where $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_n)$ is an arbitrary n-dimensional vector that satisfies the constraint

$$\sum_{i=1}^{n} |\beta_i| \leq M \ .$$

- H_n: a subset of $\mathcal{F}$, whose elements are parametrized by a number of parameters proportional to n. We will assume that the sets H_n form a nested family, that is

$$H_1 \subset H_2 \subset \ldots \subset H_n \subset \ldots \ .$$

For example H_n could be the set of polynomials in one variable of degree $n-1$, Radial Basis Functions with n centers or multilayer perceptrons with n hidden units. Notice that for Radial Basis Functions with moving centers and Multilayer perceptrons the number of parameters of an element of H_n is not n, but it is proportional to n (respectively $n(k+1)$ and $n(k+2)$, where k is the number of variables).

- H: it is defined as $H = \bigcup_{n=1}^{\infty} H_n$, and it is identified with the approximation scheme. If H_n is the set of polynomials in one variable of degree $n-1$, H is the set of polynomials of any degree.

- $H^{m,p}(R^k)$: the Sobolev space of functions in k variables whose derivatives up to order m are in $L^p(R^k)$.

- $I[f]$: the expected risk, defined as

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

where f is any function for which this expression is well defined. It is a measure of how well the function f predicts the response y.

- $I_{\text{emp}}[f]$: the empirical risk. It is a functional on $\mathcal{U}$ defined as

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

where $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ is a set of data randomly drawn from $X \times Y$ according to the probability distribution $P(\mathbf{x}, y)$. It is an approximate measure of the expected risk, since it converges to $I[f]$ in probability when the number of data points l tends to infinity.

- k: it will always indicate the number of independent variables, and therefore the dimensionality of the set X.
- l: it will always indicate the number of data points drawn from X according to the probability distribution $P(\mathbf{x})$.
- $L^2(P)$: the set of function whose square is integrable with respect to the measure defined by the probability distribution P. The norm in $L^2(P)$ is therefore defined by

$$\|f\|^2_{L^2(P)} \equiv \int_{R^k} d\mathbf{x}\ P(\mathbf{x}) f^2(\mathbf{x}) \ .$$

- $\Lambda^m(R^k)(M_0, M_1, M_2, \ldots, M_m)$: the space of functions in k variables whose derivatives up to order m are bounded:

$$|D^\alpha f| \leq M_{|\alpha|} \quad |\alpha| = 1, 2, \ldots, m$$

where α is a multi-index.

- M: a bound on the coefficients of the gaussian Radial Basis Functions technique considered in this paper, see eq. (12).
- $\mathcal{M}(\epsilon, \mathcal{S}, d)$: the packing number of the set $\mathcal{S}$, with metric d.
- $\mathcal{N}(\epsilon, \mathcal{S}, d)$: the covering number of the set $\mathcal{S}$, with metric d.
- n: a positive number proportional to the number of parameters of the approximating function. Usually will be the number of basis functions for the RBF technique or the number of hidden units for a multilayer perceptron.
- $P(\mathbf{x})$: a probability distribution defined on X. It is the probability distribution according to which the data are drawn from X.
- $P(y|\mathbf{x})$: the conditional probability of the response y given the predictor $\mathbf{x}$. It represents the probabilistic dependence of y from $\mathbf{x}$. If there is no noise in the system it has the form $P(y|\mathbf{x}) = \delta(y - h(\mathbf{x}))$, for some function h, indicating that the predictor $\mathbf{x}$ uniquely determines the response y.
- $P(\mathbf{x}, y)$: the joint distribution of the predictors and the response. It is a probability distribution on $X \times Y$ and has the form

$$P(\mathbf{x}, y) \equiv P(\mathbf{x})P(y|\mathbf{x}) \ .$$

- S: it will usually denote a metric space, endowed with a metric d.
- $\mathcal{S}$: a generic subset of a metric space S.
- $\mathcal{T}$: a generic ϵ-cover of a subset $\mathcal{S} \subset S$.
- U: it gives a bound on the elements of the class $\mathcal{A}$. In the specific case of the class $\mathcal{A}$ considere in the proof we have $U = 1 + MV$.
- $\mathcal{U}$: the set of all the functions from X to Y for which the expected risk is well defined.
- V: a bound on the Gaussian basis function G:

$$0 \leq G(\mathbf{x}) \leq V \ , \quad \forall \mathbf{x} \in R^k \ .$$

- X: a subset of R^k, not necessarily proper. It is the set of the independent variables, or predictors, or, in the language of neural networks, input variables.
- $\mathbf{x}$: a generic element of X, and therefore a k-dimensional vector (in the neural network language is the input vector).
- Y: a subset of R, whose elements represent the response variable, that in the neural networks language is the output of the network. Unless otherwise stated it will be assumed to be compact, implying that $\mathcal{F}$ is a set of bounded functions. In pattern recognition problem it is simply the set $\{0, 1\}$.
- y: a generic element of Y, it denotes the response variable.

B A Useful Decomposition of the Expected Risk

We now show that the function that minimizes the expected risk

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

is the regression function defined in eq. (3). It is sufficient to add and subtract the regression function in the definition of expected risk:

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

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

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

Since the last term of $I[f]$ does not depend on f, the minimum 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, ϵ, with probability distribution $\mathcal{P}(\epsilon)$ and zero mean, we have $P(y|\mathbf{x}) = \mathcal{P}(y - f(\mathbf{x}))$ and then

$$I[f_0] = \int_{X\times Y} d\mathbf{x}dyP(\mathbf{x},y)(y - f_0(\mathbf{x}))^2 = \quad (15)$$

$$= \int_X d\mathbf{x}P(\mathbf{x}) \int_Y (y - f(\mathbf{x}))^2 \mathcal{P}(y - f(\mathbf{x})) = \quad (16)$$

$$= \int_X d\mathbf{x}P(\mathbf{x}) \int_Y \epsilon^2 \mathcal{P}(\epsilon)d\epsilon = \sigma^2 \quad (17)$$

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.

C A Useful Inequality

Let us assume that, with probability $1 - \delta$ a uniform bound has been established:

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

We want to prove that the following inequality also holds:

$$|I[f_n] - I[\hat{f}_{n,l}]| \leq 2\omega(l, n, \delta) \ . \quad (18)$$

This fact is easily established by noting that since the bound above is uniform, then it holds for both f_n and $\hat{f}_{n,l}$, and therefore the following inequalities hold:

$$I[\hat{f}_{n,l}] \leq I_{\text{emp}}[\hat{f}_{n,l}] + \omega$$

$$I_{\text{emp}}[f_n] \leq I[f_n] + \omega$$

Moreover, by definition, the two following inequalities also hold:

$$I[f_n] \leq I[\hat{f}_{n,l}]$$

$$I_{\text{emp}}[\hat{f}_{n,l}] \leq I_{\text{emp}}[f_n]$$

Therefore tha following chain of inequalities hold, proving inequality (18):

$$I[f_n] \leq I[\hat{f}_{n,l}] \leq I_{\text{emp}}[\hat{f}_{n,l}]+\omega \leq I_{\text{emp}}[f_n]+\omega \leq I[f_n]+2\omega \ .$$

An intutitive explanation of these inequalities is also explained in figure (7).

D Proof of the Main Theorem

The theorem will be proved in a series of steps. For clarity of presentation we have divided the proof into four parts. The first takes the original problem and breaks it into its approximation and estimation components. The second and third parts are devoted to obtaining bounds for these two components respectively. The fourth and final part comes back to the original problem, reassembles its components and proves our main result. New definitions and notation will be introduced as and when the necessity arises.

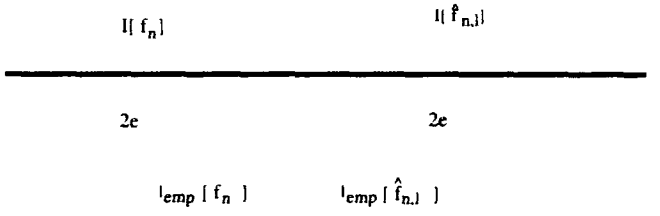

Figure 7: If the distance between $I[f_n]$ and $I[\hat{f}_{n,l}]$ is larger than 2ϵ, the condition $I_{\text{emp}}[\hat{f}_{n,l}] \leq I_{\text{emp}}[f_n]$ is violated.

We have seen in section 2 (statement 2.1) that the generalization error can be bounded, with probability $1 - \delta$, as follows:

$$\|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)} \leq \varepsilon(n) + 2\omega(l, n, \delta) \ . \quad (19)$$

In the next parts we will derive specific expressions for the approximation error ε and for the estimation error ω in order to prove theorem (4.1).

D.1 Bounding the approximation error

In this part we attempt to bound the approximation error. In section 3 we assumed that the class of functions to which the regression function belongs, that is the class of functions that we want to approximate, is

$$\mathcal{F} \equiv \{f \in L_2(R^k) | f = \lambda * G, |\lambda|_{R^k} \leq M\} \ ,$$

where λ is a signed Radon measure on the Borel sets of R^k, G is a gaussian function with range $[0, V]$, the symbol $*$ stands for the convolution operation, $|\lambda|_{R^k}$ is the total variation of the measure λ and M is a positive real number. Our approximating family is the class:

$$H_n = \{f \in L_2 | f = \sum_{i=1}^{n} \beta_i G(\mathbf{x}-\mathbf{t}_i), \sum_{i=1}^{n} |\beta_i| \leq M \ , \ \mathbf{t}_i \in R^k\}$$

It has been shown in [33, 34] that the class H_n uniformly approximate elements of $\mathcal{F}$, and that the following bound is valid:

$$E[(f_0 - f_n)^2] \leq O\left(\frac{1}{n}\right) \ . \quad (20)$$

This result is based on a lemma by Jones [48] on the convergence rate of an iterative approximation scheme in Hilbert spaces. A formally similar lemma, brought to our attention by R. Dudley [25] is due to Maurey and was published by Pisier [65]. Here we report a version

of the lemma due to Barron [6, 7] that contains a slight refinement of Jones' result:

Lemma D.1 (Maurey-Jones-Barron) *If f is in the closure of the convex hull of a set $\mathcal{G}$ in a Hilbert space H with $\|g\| \leq b$ for each $g \in \mathcal{G}$, then for every $n \geq 1$ and for $c > b^2 - \|f\|^2$ there is a f_n in the convex hull of n points in $\mathcal{G}$ such that*

$$\|f - f_n\|^2 \leq \frac{c}{n} .$$

In order to exploit this result one needs to define suitable classes of functions which are the closure of the convex hull of some subset $\mathcal{G}$ of a Hilbert space H. One way to approach the problem consists in utilizing the *integral representation* of functions. Suppose that the functions in a Hilbert space H can be represented by the integral

$$f(\mathbf{x}) = \int_{\mathcal{M}} G_{\mathbf{t}}(\mathbf{x}) d\alpha(\mathbf{t}) \qquad (21)$$

where $d\alpha$ is some measure on the parameter set $\mathcal{M}$, and $G_{\mathbf{t}}(\mathbf{x})$ is a function of H parametrized by the parameter $\mathbf{t}$, whose norm $\|G_{\mathbf{t}}(\mathbf{x})\|$ is bounded by the same number for any value of $\mathbf{t}$. If $d\alpha$ is a finite measure, the integral (21) can be seen as an infinite convex combination, and therefore, applying lemma (D.1) one can prove that there exists n coefficients c_i and n parameter vectors $\mathbf{t}_i$ such that

$$\|f - \sum_{i=1}^{n} c_i G_{\mathbf{t}_i}(\mathbf{x})\|^2 \leq O(\frac{1}{n})$$

For the class $\mathcal{F}$ we consider, it is clear that functions in this class have an integral representation of the type (21) in which $G_{\mathbf{t}}(\mathbf{x}) = G(\mathbf{x} - \mathbf{t})$, and the work in [33, 34] shows how to apply lemma (D.1) to this class.

Notice that the bound (20), that is similar in spirit to the result of A. Barron on multilayer perceptrons [6, 8], is interesting because the rate of convergence does not depend on the dimension d of the input space. This is apparently unusual in approximation theory, because it is known, from the theory of linear and nonlinear widths [78, 64, 54, 55, 20, 19, 21, 56], that, if the function that has to be approximated has d variables and a degree of smoothness s, we should not expect to find an approximation technique whose approximation error goes to zero faster than $O(n^{-\frac{s}{d}})$. Here "degree of smoothness" is a measure of how constrained the class of functions we consider is, for example the number of derivatives that are uniformly bounded, or the number of derivatives that are integrable or square integrable. Therefore, from classical approximation theory, we expect that, *unless certain constraints are imposed on the class of functions to be approximated*, the rate of convergence will dramatically slow down as the number of dimensions increases, showing the phenomenon known as "the curse of dimensionality" [11].

In the case of class $\mathcal{F}$ we consider here, the constraint of considering functions that are convolutions of Radon measures with Gaussian seems to impose on this class of functions an amount of smoothness that is sufficient to guarantee that the rate of convergence does not become slower and slower as the dimension increases. A longer discussion of the "curse of dimensionality" can be found in [34].

We notice also that, since the rate (20) is independent of the dimension, the class $\mathcal{F}$, together with the approximating class H_n, defines a class of problems that are "tractable" even in a high number of dimensions.

D.2 Bounding the estimation error

In this part we attempt to bound the estimation error $|I[f] - I_{\text{emp}}[f]|$. In order to do that we first need to introduce some basic concepts and notations.

Let $\mathcal{S}$ be a subset of a metric space S with metric d. We say that an **ϵ-cover** with respect to the metric d is a set $\mathcal{T} \in S$ such that for every $s \in \mathcal{S}$, there exists some $t \in \mathcal{T}$ satisfying $d(s,t) \leq \epsilon$. The size of the smallest ϵ-cover is $\mathcal{N}(\epsilon, \mathcal{S}, d)$ and is called the **covering number** of $\mathcal{S}$. In other words

$$\mathcal{N}(\epsilon, \mathcal{S}, d) = \min_{\mathcal{T} \subset \mathcal{S}} |\mathcal{T}| ,$$

where $\mathcal{T}$ runs over all the possible ϵ-cover of $\mathcal{S}$ and $|\mathcal{T}|$ denotes the cardinality of $\mathcal{T}$.

A set B belonging to the metric space S is said to be **ϵ-separated** if for all $x, y \in B$, $d(x,y) > \epsilon$. We define the the *packing number* $\mathcal{M}(\epsilon, \mathcal{S}, d)$ as the size of the largest ϵ-separated subset of $\mathcal{S}$. Thus

$$\mathcal{M}(\epsilon, \mathcal{S}, d) = \max_{B \subset \mathcal{S}} |B| ,$$

where B runs over all the ϵ-separated subsets of $\mathcal{S}$. It is easy to show that the covering number is always less than the packing number, that is $\mathcal{N}(\epsilon, \mathcal{S}, d) \leq \mathcal{M}(\epsilon, \mathcal{S}, d)$.

Let now $P(\xi)$ be a probability distribution defined on S, and $\mathcal{A}$ be a set of real-valued functions defined on S such that, for any $a \in \mathcal{A}$,

$$0 \leq a(\xi) \leq U^2 \quad \forall \xi \in S .$$

Let also $\bar{\xi} = (\xi_1, .., \xi_l)$ be a sequence of l examples drawn independently from S according to $P(\xi)$. For any function $a \in \mathcal{A}$ we define the empirical and true expectations of a as follows:

$$\hat{E}[a] = \frac{1}{l} \sum_{i=1}^{l} a(\xi_i)$$

$$E[a] = \int_S d\xi P(\xi) a(\xi)$$

The difference between the empirical and true expectation can be bounded by the following inequality, whose proof can be found in [69] and [42], that will be crucial in order to prove our main theorem.

Claim D.1 ([69], [42]) *Let $\mathcal{A}$ and $\bar{\xi}$ be as defined above. Then, for all $\epsilon > 0$,*

$$P\left(\exists a \in \mathcal{A} : |\hat{E}[a] - E[a]| > \epsilon\right) \leq$$

$$\leq 4E\left[\mathcal{N}(\tfrac{\epsilon}{16}, \mathcal{A}_{\bar{\xi}}, d_{L^1})\right] e^{-\frac{1}{128U^4}\epsilon^2 l}$$

In the above result, $\mathcal{A}_{\bar{\xi}}$ is the restriction of $\mathcal{A}$ to the data set, that is:

$$\mathcal{A}_{\bar{\xi}} \equiv \{(a(\xi_1), \ldots, a(\xi_l)) : a \in \mathcal{A}\} . \qquad (22)$$

The set $\mathcal{A}_{\bar{\xi}}$ is a collection of points belonging to the subset $[0, U]^l$ of the l-dimensional euclidean space. Each function a in $\mathcal{A}$ is represented by a point in $\mathcal{A}_{\bar{\xi}}$, while every point in $\mathcal{A}_{\bar{\xi}}$ represents all the functions that have the same values at the points $\xi_1, \ldots, \xi_l$. The distance metric d_{L^1} in the inequality above is the standard L^1 metric in R^l, that is

$$d_{L^1}(\mathbf{x}, \mathbf{y}) = \frac{1}{l} \sum_{\mu=1}^{l} |x^\mu - y^\mu|$$

where $\mathbf{x}$ and $\mathbf{y}$ are points in the l-dimensional euclidean space and x^μ and y^μ are their μ-th components respectively.

The above inequality is a result in the theory of uniform convergence of empirical measures to their underlying probabilities, that has been studied in great detail by Pollard and Vapnik, and similar inequalities can be found in the work of Vapnik [81, 82, 80], although they usually involve the VC dimension of the set $\mathcal{A}$, rather than its covering numbers.

Suppose now we choose $S = X \times Y$, where X is an arbitrary subset of R^k and $Y = [-M, M]$ as in the formulation of our original problem. The generic element of S will be written as $\xi = (\mathbf{x}, y) \in X \times Y$. We now consider the class of functions $\mathcal{A}$ defined as:

$$\mathcal{A} = \{a : X \times Y \to R \mid a(\mathbf{x}, y) = (y - h(\mathbf{x}))^2,\ h \in H_n(R^k)\}$$

where $H_n(R^k)$ is the class of k-dimensional Radial Basis Functions with n basis functions defined in eq. 12 in section 3. Clearly,

$$|y - h(\mathbf{x})| \leq |y| + |h(\mathbf{x})| \leq M + MV,$$

and therefore

$$0 \leq a \leq U^2$$

where we have defined

$$U \equiv M + MV .$$

We notice that, by definition of $\hat{E}(a)$ and $E(a)$ we have

$$\hat{E}(a) = \frac{1}{l} \sum_{i=1}^{l} (y_i - h(\mathbf{x}_i))^2 = I_{\text{emp}}[h]$$

and

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

Therefore, applying the inequality of claim D.1 to the set $\mathcal{A}$, and noticing that the elements of $\mathcal{A}$ are essentially defined by the elements of H_n, we obtain the following result:

$$P(\forall h \in H_n, |I_{\text{emp}}[h] - I[h]| \leq \epsilon) \geq$$
$$\geq 1 - 4E[\mathcal{N}(\epsilon/16, \mathcal{A}_{\bar{\xi}}, d_{L^1})] e^{-\frac{1}{128U^4}\epsilon^2 l} . \qquad (23)$$

so that the inequality of claim D.1 gives us a bound on the estimation error. However, this bound depends on the specific choice of the probability distribution $P(\mathbf{x}, y)$, while we are interested in bounds that do not depend on P. Therefore it is useful to define some quantity that does not depend on P, and give bounds in terms of that.

We then introduce the concept of **metric capacity** of $\mathcal{A}$, that is defined as

$$\mathcal{C}(\epsilon, \mathcal{A}, d_{L^1}) = \sup_P \{\mathcal{N}(\epsilon, \mathcal{A}, d_{L^1(P)})\}$$

where the supremum is taken over all the probability distributions P defined over S, and $d_{L^1(P)}$ is standard $L^1(P)$ distance[9]
induced by the probability distribution P:

$$d_{L^1(P)}(a_1, a_2) = \int_S d\xi P(\xi) |a_1(\xi) - a_2(\xi)| \quad a_1, a_2 \in \mathcal{A} .$$

The relationship between the covering number and the metric capacity is showed in the following

Claim D.2

$$E[\mathcal{N}(\epsilon, \mathcal{A}_{\bar{\xi}}, d_{L^1})] \leq \mathcal{C}(\epsilon, \mathcal{A}, d_{L^1}) .$$

Proof: For any sequence of points $\bar{\xi}$ in S, there is a trivial isometry between $(\mathcal{A}_{\bar{\xi}}, d_{L^1})$ and $(\mathcal{A}, d_{L^1(P_{\bar{\xi}})})$ where $P_{\bar{\xi}}$ is the empirical distribution on the space S given by $\frac{1}{l}\sum_{i=1}^{l} \delta(\xi - \xi_i)$. Here δ is the Dirac delta function, $\xi \in S$, and ξ_i is the i-th element of the data set. To see that this isometry exists, first note that for every element $a \in \mathcal{A}$, there exists a unique point $(a(\xi_1), \ldots, a(\xi_l)) \in \mathcal{A}_{\bar{\xi}}$. Thus a simple bijective mapping exists between the two spaces. Now consider any two elements g and h of $\mathcal{A}$. The distance between them is given by

$$d_{L^1(P_{\bar{\xi}})}(g, h) = \int_S |g(\xi) - h(\xi)| P_{\bar{\xi}}(\xi) d\xi = \frac{1}{l} \sum_{i=1}^{l} |g(\xi_i) - h(\xi_i)| .$$

This is exactly what the distance between the two points $(g(\xi_1), .., g(\xi_l))$ and $(h(\xi_1), .., h(\xi_l))$, which are elements of $\mathcal{A}_{\bar{\xi}}$, is according to the d_{L^1} distance. Thus there is

[9] Note that here $\mathcal{A}$ is a class of real-valued functions defined on a general metric space S. If we consider an arbitrary $\mathcal{A}$ defined on S and taking values in R^n, the $d_{L^1(P)}$, norm is appropriately adjusted to be

$$d_{L^1(P)}(\mathbf{f}, \mathbf{g}) = \frac{1}{n} \sum_{i=1}^{n} \int_S |f_i(\mathbf{x}) - g_i(\mathbf{x})| P(\mathbf{x}) d\mathbf{x}$$

where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots f_i(\mathbf{x}), \ldots f_n(\mathbf{x}))$, $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots g_i(\mathbf{x}), \ldots g_n(\mathbf{x}))$ are elements of $\mathcal{A}$ and $P(\mathbf{x})$ is a probability distribution on S. Thus d_{L^1} and $d_{L^1(P)}$ should be interpreted according to the context.

a one-to-one correspondence between elements of $\mathcal{A}$ and $\mathcal{A}_{\tilde{\xi}}$ and the distance between two elements in $\mathcal{A}$ is the same as the distance between their corresponding points in $\mathcal{A}_{\tilde{\xi}}$. Given this isometry, for every ϵ-cover in $\mathcal{A}$, there exists an ϵ-cover of the same size in $\mathcal{A}_{\tilde{\xi}}$, so that

$$\mathcal{N}(\epsilon, \mathcal{A}_{\tilde{\xi}}, d_{L^1}) = \mathcal{N}(\epsilon, \mathcal{A}, d_{L^1(P_\xi)}) \leq \mathcal{C}(\epsilon, \mathcal{A}, d_{L^1}).$$

and consequently $E[\mathcal{N}(\epsilon, \mathcal{A}_{\tilde{\xi}}, d_{L^1})] \leq \mathcal{C}(\epsilon, \mathcal{A}, d_{L^1})$. □

The result above, together with eq. (23) shows that the following proposition holds:

Claim D.3

$$P(\forall h \in H_n, |I_{\text{emp}}[h] - I[h]| \leq \epsilon) \geq \geq 1 - 4\mathcal{C}(\epsilon/16, \mathcal{A}, d_{L^1})]e^{-\frac{1}{128U^4}\epsilon^2 l} . \quad (24)$$

Thus in order to obtain a uniform bound ω on $|I_{\text{emp}}[h] - I[h]|$, our task is reduced to computing the metric capacity of the functional class $\mathcal{A}$ which we have just defined. We will do this in several steps. In Claim D.4, we first relate the metric capacity of $\mathcal{A}$ to that of the class of radial basis functions H_n. Then Claims D.5 through D.9 go through a computation of the metric capacity of H_n.

Claim D.4

$$C(\epsilon, \mathcal{A}, d_{L^1}) \leq C(\epsilon/4U, H_n, d_{L^1})$$

Proof: Fix a distribution P on $S = X \times Y$. Let P_X be the marginal distribution with respect to X. Suppose K is an $\epsilon/4U$-cover for H_n with respect to this probability distribution P_X, i.e. with respect to the distance metric $d_{L^1(P_X)}$ on H_n. Further let the size of K be $\mathcal{N}(\epsilon/4U, H_n, d_{L^1(P_X)})$. This means that for any $h \in H_n$, there exists a function h^* belonging to K, such that:

$$\int |h(\mathbf{x}) - h^*(\mathbf{x})| P_X(\mathbf{x}) d\mathbf{x} \leq \epsilon/4U$$

Now we claim the set $H(K) = \{(y - h(\mathbf{x}))^2 : h \in K\}$ is an ϵ cover for $\mathcal{A}$ with respect to the distance metric $d_{L^1(P)}$. To see this, it is sufficient to show that

$$\int |(y - h(\mathbf{x}))^2 - (y - h^*(\mathbf{x}))^2| P(\mathbf{x}, y) d\mathbf{x} dy \leq$$

$$\leq \int 2|(2y - h - h^*)||(h - h^*)| P(\mathbf{x}, y) d\mathbf{x} dy \leq$$

$$\leq \int 2(2M + 2MV)|h - h^*| P(\mathbf{x}, y) d\mathbf{x} dy \leq \epsilon$$

which is clearly true. Now

$$\mathcal{N}(\epsilon, \mathcal{A}, d_{L^1(P)}) \leq |H(K)| =$$

$$= \; calN(\epsilon/4U, H_n, d_{L^1(P_X)}) \leq$$

$$\leq \mathcal{C}(\epsilon/4U, H_n, d_{L^1})$$

Taking the supremum over all probability distributions, the result follows. □

So the problem reduces to finding $C(\epsilon, H_n, d_{L^1})$, i.e. the metric capacity of the class of appropriately defined Radial Basis Functions networks with n centers. To do this we will decompose the class H_n to be the composition of two classes defined as follows.

Definitions/Notations

H_I is a class of functions defined from the metric space (R^k, d_{L^1}) to the metric space (R^n, d_{L^1}). In particular,

$$H_I = \{\mathbf{g}(\mathbf{x}) = (G(||\mathbf{x}-\mathbf{t}_1||), G(||\mathbf{x}-\mathbf{t}_2||), \ldots, G(||\mathbf{x}-\mathbf{t}_n||))\}$$

where G is a Gaussian and $\mathbf{t}_i$ are k-dimensional vectors. Note here that G is the same Gaussian that we have been using to build our Radial-Basis-Function Network. Thus H_I is parametrized by the n centers t_i and the variance of the Gaussian σ^2, in other words $nk + 1$ parameters in all.

H_F is a class defined from the metric space $([0, V]^n, d_{L^1})$ to the metric space (R, d_{L^1}). In particular,

$$H_F = \{h(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x}, \;\; \mathbf{x} \in [0, V]^n \text{ and } \sum_{i=1}^{n} |\beta_i| \leq M\}$$

where $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_n)$ is an arbitrary n-dimensional vector.

Thus we see that

$$H_n = \{h_F \circ h_I : h_F \in H_F \text{ and } h_I \in H_I\}$$

where $\circ$ stands for the composition operation, i.e., for any two functions f and g, $f \circ g = f(g(\mathbf{x}))$. It should be pointed out that H_n as defined above is defined from R^k to R.

Claim D.5

$$C(\epsilon, H_I, d_{L^1}) \leq 2^n \left(\frac{2eV}{\epsilon} \ln\left(\frac{2eV}{\epsilon}\right)\right)^{n(k+2)}$$

Proof: Fix a probability distribution P on R^k. Consider the class

$$\mathcal{G} = \{g : g(\mathbf{x}) = G(||\mathbf{x} - \mathbf{t}||), \; \mathbf{t} \in R^k\}.$$

Let K be an $\mathcal{N}(\epsilon, \mathcal{G}, d_{L^1(P)})$-sized ϵ cover for this class. We first claim that

$$T = \{(h_1, .., h_n) : h_i \in K\}$$

is an ϵ-cover for H_I with respect to the $d_{L^1(P)}$ metric.

Remember that the $d_{L^1(P)}$ distance between two vector-valued functions $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), .., g_n(\mathbf{x}))$ and $\mathbf{g}^*(\mathbf{x}) = (g_1^*(\mathbf{x}), .., g_n^*(\mathbf{x}))$ is defined as

$$d_{L^1(P)}(\mathbf{g}, \mathbf{g}^*) = \frac{1}{n} \sum_{i=1}^{n} \int |g_i(\mathbf{x}) - g_i^*(\mathbf{x})| P(\mathbf{x}) d\mathbf{x}$$

To see this, pick an arbitrary $\mathbf{g} = (g_1, \ldots, g_n) \in H_I$. For each g_i, there exists a $g_i^* \in K$ which is ϵ-close

in the appropriate sense for real-valued functions, i.e. $d_{L^1(P)}(g_i, g_i^*) \leq \epsilon$. The function $\mathbf{g} = (g_1^*, .., g_n^*)$ is an element of T. Also, the distance between $(g_1, .., g_n)$ and $(g_1^*, .., g_n^*)$ in the $d_{L^1(P)}$ metric is

$$d_{L^1(P)}(\mathbf{g}, \mathbf{g}^*) \leq \frac{1}{n} \sum_{i=1}^{n} \epsilon = \epsilon .$$

Thus we obtain that

$$\mathcal{N}(\epsilon, H_I, d_{L^1(P)}) \leq [\mathcal{N}(\epsilon, \mathcal{G}, d_{L^1(P)})]^n$$

and taking the supremum over all probability distributions as usual, we get

$$\mathcal{C}(\epsilon, H_I, d_{L^1}) \leq (\mathcal{C}(\epsilon, \mathcal{G}, d_{L^1}))^n .$$

Now we need to find the capacity of $\mathcal{G}$. This is done in the Claim D.6. From this the result follows. □

Definitions/Notations

Before we proceed to the next step in our proof, some more notation needs to be defined. Let $\mathcal{A}$ be a family of functions from a set S into R. For any sequence $\bar{\xi} = (\xi_1, .., \xi_d)$ of points in S, let $\mathcal{A}_{\bar{\xi}}$ be the restriction of $\mathcal{F}$ to the data set, as per our previously introduced notation. Thus $\mathcal{A}_{\bar{\xi}} = \{(a(\xi_1), \ldots, a(\xi_d)) : a \in \mathcal{A}\}$. If there exists some translation of the set $\mathcal{A}_{\bar{\xi}}$, such that it intersects all 2^d orthants of the space R^d, then $\bar{\xi}$ is said to be **shattered** by $\mathcal{A}$. Expressing this a little more formally, let $\mathcal{B}$ be the set of all possible l-dimensional boolean vectors. If there exists a translation $\mathbf{t} \in R^d$ such that for every $\mathbf{b} \in \mathcal{B}$, there exists some function $a_{\mathbf{b}} \in \mathcal{A}$ satisfying $a_{\mathbf{b}}(\xi_i) - t_i \geq b_i \Leftrightarrow b_i = 1$ for all $i = 1$ to d, then the set $(\xi_1, .., \xi_d)$ is shattered by $\mathcal{A}$. Note that the inequality could easily have been defined to be strict and would not have made a difference. The largest d such that there exists a sequence of d points which are shattered by $\mathcal{A}$ is said to be the pseudo-dimension of $\mathcal{A}$ denoted by pdim$\mathcal{A}$. □

In this context, there are two important theorems which we will need to use. We give these theorems without proof.

Theorem D.1 (Dudley) *Let F be a k-dimensional vector space of functions from a set S into R. Then pdim$(F) = k$.*

The following theorem is stated and proved in a somewhat more general form by Pollard. Haussler, using techniques from Pollard has proved the specific form shown here.

Theorem D.2 (Pollard, Haussler) *Let F be a family of functions from a set S into $[M_1, M_2]$, where pdim$(F) = d$ for some $1 \leq d < \infty$. Let P be a probability distribution on S. Then for all $0 < \epsilon \leq M_2 - M_1$,*

$$\mathcal{M}(\epsilon, F, d_{L^1(P)}) < 2 \left(\frac{1}{\epsilon} 2e(M_2 - M_1) \log \frac{1}{\epsilon} 2e(M_2 - M_1) \right)^d$$

Here $\mathcal{M}(\epsilon, F, d_{L^1(P)})$ is the packing number of F according to the distance metric $d_{L^1(P)}$.

Claim D.6

$$C(\epsilon, \mathcal{G}, d_{L^1}) \leq 2 \left(\frac{2eV}{\epsilon} \ln \left(\frac{2eV}{\epsilon} \right) \right)^{(k+2)}$$

Proof: Consider the $k+2$-dimensional vector space of functions from R^k to R defined as

$$G_1 \equiv \text{span}\{1, x^1, x^2, \cdot, x^k, ||\mathbf{x}||^2\}$$

where $\mathbf{x} \in R^k$ and x^μ is the μ-th component of the vector $\mathbf{x}$. Now consider the class

$$G_2 = \{\alpha e^{-f} : f \in G_1, \quad \alpha = \frac{1}{\sqrt{2\pi}\sigma}\}$$

where σ is the standard deviation of the Gaussian, and $f \in G_1$. We claim that the pseudo-dimension of $\mathcal{G}$ denoted by pdim$(\mathcal{G})$ fulfills the following inequality,

$$\text{pdim}\ (\mathcal{G}) \leq \ \text{pdim}\ (G_2) = \ \text{pdim}\ (G_1) = (k+2).$$

To see this consider the fact that $\mathcal{G} \subset G_2$. Consequently, for every sequence of points $\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$, $\mathcal{G}_{\bar{\mathbf{x}}} \subset (G_2)_{\bar{\mathbf{x}}}$. Thus if $(\mathbf{x}_1, \ldots, \mathbf{x}_d)$ is shattered by $\mathcal{G}$, it will be shattered by G_2. This establishes the first inequality.

We now show that pdim$(G_2) \leq$ pdim(G_1). It is enough to show that every set shattered by G_2 is also shattered by G_1. Suppose there exists a sequence $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)$ which is shattered by G_2. This means that by our definition of shattering, there exists a translation $\mathbf{t} \in R^d$ such that for every boolean vector $\mathbf{b} \in \{0,1\}^d$ there is some function $g_{\mathbf{b}} = \alpha e^{-f_{\mathbf{b}}}$ where $f_{\mathbf{b}} \in G_1$ satisfying $g_{\mathbf{b}}(x_i) \geq t_i$ if and only if $b_i = 1$, where t_i and b_i are the i-th components of $\mathbf{t}$ and $\mathbf{b}$ respectively. First notice that every function in G_2 is positive. Consequently, we see that every t_i has to be greater than 0, for otherwise, $g_b(\mathbf{x}_i)$ could never be less than t_i which it is required to be if $b_i = 0$. Having established that every t_i is greater than 0, we now show that the set $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)$ is shattered by G_1. We let the translation in this case be $\mathbf{t}' = (\log(t_1/\alpha), \log(t_2/\alpha), \ldots, \log(t_d/\alpha))$. We can take the log since the t_i/α's are greater than 0. Now for every boolean vector $\mathbf{b}$, we take the function $-f_b \in G_1$ and we see that since

$$g_b = \alpha e^{-f_b} \geq t_i \Leftrightarrow b_i = 1.$$

if follows that

$$-f_b \geq \log(t_i/\alpha) = \mathbf{t}'_i \Leftrightarrow b_i = 1.$$

Thus we see that the set $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)$ can be shattered by G_1. By a similar argument, it is also possible to show that pdim$(G_1) \geq$ pdim(G_2).

Since G_1 is a vector space of dimensionality $k+2$, an application of Dudley's Theorem [24] yields the value $k+2$ for its pseudo-dimension. Further, functions in the class $\mathcal{G}$ are in the range $[0, V]$. Now we see (by an application of Pollard's theorem) that

$$\mathcal{N}(\epsilon, \mathcal{G}, d_{L^1(P)}) \leq \mathcal{M}(\epsilon, \mathcal{G}, d_{L^1(P)}) \leq$$

$$\leq 2\left(\tfrac{2eV}{\epsilon} \ln\left(\tfrac{2eV}{\epsilon}\right)\right)^{\text{pdim}(\mathcal{G})} \leq$$

$$\leq 2\left(\tfrac{2eV}{\epsilon} \ln\left(\tfrac{2eV}{\epsilon}\right)\right)^{(k+2)}$$

Taking the supremum over all probability distributions, the result follows.□

Claim D.7

$$C(\epsilon, H_F, d_{L^1}) \leq 2\left(\frac{4MeV}{\epsilon} \ln\left(\frac{4MeV}{\epsilon}\right)\right)^n$$

Proof: The proof of this runs in very similar fashion. First note that

$$H_F \subset \{\boldsymbol{\beta} \cdot \mathbf{x} : \mathbf{x},\ \boldsymbol{\beta} \in R^n\}.$$

The latter set is a vector space of dimensionality n and by Dudley's theorem[24], we see that its pseudo-dimension pdim is n. Also, clearly by the same argument as in the previous proposition, we have that $\text{pdim}(H_F) \leq n$. To get bounds on the functions in H_F, notice that

$$|\sum_{i=1}^{n} \beta_i x_i| \leq \sum_{i=1}^{n} |\beta_i||x_i| \leq V \sum_{i=1}^{n} |\beta_i| \leq MV.$$

Thus functions in H_F are bounded in the range $[-MV, MV]$. Now using Pollard's result [42], [69], we have that

$$\mathcal{N}(\epsilon, H_F, d_{L^1(P)}) \leq \mathcal{M}(\epsilon, H_F, d_{L^1(P)}) \leq$$

$$\leq 2\left(\tfrac{4MeV}{\epsilon} \ln\left(\tfrac{4MeV}{\epsilon}\right)\right)^n.$$

Taking supremums over all probability distributions, the result follows. □

Claim D.8 *A uniform first-order Lipschitz bound of H_F is Mn.*

Proof: Suppose we have $\mathbf{x},\ \mathbf{y} \in R^n$ such that

$$d_{L^1}(\mathbf{x}, \mathbf{y}) \leq \epsilon.$$

The quantity Mn is a uniform first-order Lipschitz bound for H_F if, for any element of H_F, parametrized by a vector $\boldsymbol{\beta}$, the following inequality holds:

$$|\mathbf{x} \cdot \boldsymbol{\beta} - \mathbf{y} \cdot \boldsymbol{\beta}| \leq Mn\epsilon$$

Now clearly,

$$|\mathbf{x} \cdot \boldsymbol{\beta} - \mathbf{y} \cdot \boldsymbol{\beta}| = |\textstyle\sum_{i=1}^{n} \beta_i(x_i - y_i)| \leq$$

$$\leq \textstyle\sum_{i=1}^{n} |\beta_i||(x_i - y_i)| \leq$$

$$\leq M \textstyle\sum_{i=1}^{n} |(x_i - y_i)| \leq Mn\epsilon$$

The result is proved. □

Claim D.9

$$C(\epsilon, H_n, d_{L^1}) \leq C(\frac{\epsilon}{2Mn}, H_I, d_{L^1}) C(\frac{\epsilon}{2}, H_F, d_{L^1})$$

Proof: Fix a distribution P on R^k. Assume we have an $\epsilon/(2Mn)$-cover for H_I with respect to the probability distribution P and metric $d_{L^1(P)}$. Let it be K where

$$|K| = \mathcal{N}(\epsilon/2Mn, H_I, d_{L^1(P)}).$$

Now each function $f \in K$ maps the space R^k into R^n, thus inducing a probability distribution P_f on the space R^n. Specifically, P_f can be defined as the distribution obtained from the measure μ_f defined so that any measurable set $A \subset R^n$ will have measure

$$\mu_f(A) = \int_{f^{-1}(A)} P(\mathbf{x})d\mathbf{x}\ .$$

Further, there exists a cover K_f which is an $\epsilon/2$-cover for H_F with respect to the probability distribution P_f. In other words

$$|K_f| = \mathcal{N}(\epsilon/2, H_F, d_{L^1(P_f)}).$$

We claim that

$$H(K) = \{f \circ g : g \in K \text{ and } f \in K_g\}$$

is an ϵ cover for H_n. Further we note that

$$|H(K)| = \textstyle\sum_{f \in K} |K_f| \leq \sum_{f \in K} \mathcal{C}(\epsilon/2, H_F, d_{L^1}) \leq$$

$$\leq \mathcal{N}(\epsilon/(2Mn), H_I, d_{L^1(P)}) \mathcal{C}(\epsilon/2, H_F, d_{L^1})$$

To see that $H(K)$ is an ϵ-cover, suppose we are given an arbitrary function $h_f \circ h_i \in H_n$. There clearly exists a function $h_i^* \in K$ such that

$$\int_{R^k} d_{L^1}(h_i(\mathbf{x}), h_i^*(\mathbf{x}))P(\mathbf{x})d\mathbf{x} \leq \epsilon/(2Mn)$$

Now there also exists a function $h_f^* \in K_{h_i^*}$ such that

$$\textstyle\int_{R^k} |h_f \circ h_i^*(\mathbf{x}) - h_f^* \circ h_i^*(\mathbf{x})|P(\mathbf{x})d\mathbf{x} =$$

$$= \textstyle\int_{R^n} |h_f(\mathbf{y}) - h_f^*(\mathbf{y})|P_{h_i^*}(\mathbf{y})d\mathbf{y} \leq \epsilon/2\ .$$

To show that $H(K)$ is an ϵ-cover it is sufficient to show that

$$\int_{R^k} |h_f \circ h_i(\mathbf{x}) - h_f^* \circ h_i^*(\mathbf{x})|P(\mathbf{x})d\mathbf{x} \leq \epsilon.$$

Now

$$\textstyle\int_{R^k} |h_f \circ h_i(\mathbf{x}) - h_f^* \circ h_i^*(\mathbf{x})|P(\mathbf{x})d\mathbf{x} \leq$$

$$\leq \textstyle\int_{R^k} \{|h_f \circ h_i(\mathbf{x}) - h_f \circ h_i^*(\mathbf{x})| +$$

$$+ |h_f \circ h_i^*(\mathbf{x}) - h_f^* \circ h_i^*(\mathbf{x})|P(\mathbf{x})d\mathbf{x}\}$$

by the triangle inequality. Further, since h_f is Lipschitz bounded,

$$\int_{R^k} |h_f \circ h_i(\mathbf{x}) - h_f \circ h_i^*(\mathbf{x})| P(\mathbf{x}) d\mathbf{x} \leq$$

$$\leq \int_{R^k} M n d_{L^1}(h_i(\mathbf{x}), h_i^*(\mathbf{x})) P(\mathbf{x}) d\mathbf{x} \leq Mn(\epsilon/2Mn) \leq \epsilon/2 .$$

Also,

$$\int_{R^k} |h_f \circ h_i^*(\mathbf{x}) - h_f^* \circ h_i^*(\mathbf{x})| P(\mathbf{x}) d\mathbf{x} =$$

$$= \int_{R^n} |h_f(\mathbf{y}) - h_f^*(\mathbf{y})| P_{h_i^*}(\mathbf{y}) d\mathbf{y} \leq \epsilon/2 .$$

Consequently both sums are less than $\epsilon/2$ and the total integral is less than ϵ. Now we see that

$$\mathcal{N}(\epsilon, H_n, d_{L^1(P)}) \leq \mathcal{N}\left(\epsilon/(2Mn), H_I, d_{L^1(P)}\right) C(\epsilon/2, H_F, d_{L^1}).$$

Taking supremums over all probability distributions, the result follows. □

Having obtained the crucial bound on the metric capacity of the class H_n, we can now prove the following

Claim D.10 *With probability $1-\delta$, and $\forall h \in H_n$, the following bound holds:*

$$|I_{\text{emp}}[h] - I[h]| \leq O\left(\left[\frac{nk\ln(nl) + \ln(1/\delta)}{l}\right]^{1/2}\right)$$

Proof: We know from the previous claim that

$$C(\epsilon, H_n, d_{L^1}) \leq$$

$$\leq 2^{n+1}\left[\tfrac{4MeVn}{\epsilon}\ln\left(\tfrac{4MeVn}{\epsilon}\right)\right]^{n(k+2)}\left[\tfrac{8MeV}{\epsilon}\ln\left(\tfrac{8MeV}{\epsilon}\right)\right]^{n} \leq$$

$$\leq \left[\tfrac{8MeVn}{\epsilon}\ln\left(\tfrac{8MeVn}{\epsilon}\right)\right]^{n(k+3)} .$$

From claim (D.3), we see that

$$P(\forall h \in H_n, |I_{\text{emp}}[h] - I[h]| \leq \epsilon) \geq \quad (25)$$

$$\geq 1 - \delta$$

as long as

$$C(\epsilon/16, \mathcal{A}, d_{L^1}) e^{-\frac{1}{128U^4}\epsilon^2 l} \leq \frac{\delta}{4}$$

which in turn is satisfied as long as (by Claim D.4)

$$C(\epsilon/64U, H_n, d_{L^1}) e^{-\frac{1}{128U^2}\epsilon^2 l} \leq \frac{\delta}{4}$$

which implies

$$\left(\tfrac{1}{\epsilon}256MeVUn\ \ln\left(\tfrac{1}{\epsilon}256MeVUn\right)\right)^{n(k+3)} \cdot$$

$$\cdot e^{-\frac{1}{128U^2}\epsilon^2 l} \leq \tfrac{\delta}{4}$$

In other words,

$$\left(\frac{An}{\epsilon}\ln\left(\frac{An}{\epsilon}\right)\right)^{n(k+3)} e^{-\epsilon^2 l/B} \leq \frac{\delta}{4}$$

for constants A, B. The latter inequality is satisfied as long as

$$(An/\epsilon)^{2n(k+3)} e^{-\epsilon^2 l/B} \leq \frac{\delta}{4}$$

which implies

$$2n(k+3)(\ln(An) - \ln(\epsilon)) - \epsilon^2 l/B \leq \ln(\delta/4)$$

and in turn implies

$$\epsilon^2 l > B\ln(4/\delta) + 2Bn(k+3)(\ln(An) - \ln(\epsilon)).$$

We now show that the above inequality is satisfied for

$$\epsilon = \left(\frac{B\left[\ln(4/\delta) + 2n(k+3)\ln(An) + n(k+3)\ln(l)\right]}{l}\right)^{1/2}$$

Putting the above value of ϵ in the inequality of interest, we get

$$\epsilon^2(l/B) = \ln(4/\delta) + 2n(k+3)\ln(An) + n(k+3)\ln(l) \geq$$

$$\geq \ln(4/\delta) + 2n(k+3)\ln(An) +$$

$$+2n(k+3)\tfrac{1}{2}\ln\left(\tfrac{l}{B[\ln(4/\delta)+2n(k+3)\ln(An)+n(k+3)\ln(l)]}\right)$$

In other words,

$$n(k+3)\ln(l) \geq$$

$$\geq n(k+3)\ln\left(\tfrac{l}{B[\ln(4/\delta)+2n(k+3)\ln(An)+n(k+3)\ln(l)]}\right)$$

Since

$$B\left[\ln(4/\delta) + 2n(k+3)\ln(An) + n(k+3)\ln(l)\right] \geq 1$$

the inequality is obviously true for this value of ϵ. Taking this value of ϵ then proves our claim. □

D.3 Bounding the generalization error

Finally we are able to take our results in Parts II and III to prove our main result:

Theorem D.3 *With probability greater than $1-\delta$ the following inequality is valid:*

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

Proof: We have seen in statement (2.1) that the generalization error is bounded as follows:

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

In section (D.1) we showed that

$$\varepsilon(n) = O\left(\frac{1}{n}\right)$$

and in claim (D.10) we showed that

$$\omega(l, n, \delta) = O\left(\left[\frac{nk\ln(nl) - \ln\delta}{l}\right]^{1/2}\right) .$$

Therefore the theorem is proved putting these results together. □

References

[1] D. E. Rumelhart A. S. Weigand 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.

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

[3] W. Arai. Mapping abilities of three-layer networks. In *Proceedings of the International Joint Conference on Neural Networks*, pages I–419–I–423, Washington D.C., June 1989. IEEE TAB Neural Network Committee.

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

[5] A.R. Barron. Approximation and estimation bounds for artificial neural networks. Technical Report 59, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, March 1991.

[6] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. Technical Report 58, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, March 1991.

[7] A.R. Barron. Approximation and estimation bounds for artificial neural networks. Technical Report 59, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, March 1991a.

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

[9] E. B. Baum and D. Haussler. What size net gives valid generalization? In *IEEE Int. Symp. Inform. Theory*, Kobe, Japan, June 1988.

[10] E. B. Baum and D. Haussler. What size net gives valid generalization? In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems I*, pages 81–90. Morgan Kaufmann Publishers, Carnegie Mellon University, 1989.

[11] R.E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

[12] 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.

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

[14] D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[15] C. K. Chui and X. Li. Approximation by ridge functions and neural networks with one hidden layer. CAT Report 222, Texas A and M University, 1990.

[16] 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.

[17] 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.

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

[19] R. DeVore, R. Howard, and C. Micchelli. Optimal nonlinear approximation. *Manuskripta Mathematika*, 1989.

[20] R.A. DeVore. Degree of nonlinear approximation. In C.K. Chui, L.L. Schumaker, and D.J. Ward, editors, *Approximation Theory, VI*, pages 175–201. Academic Press, New York, 1991.

[21] R.A. DeVore and X.M. Yu. Nonlinear n-widths in Besov spaces. In C.K. Chui, L.L. Schumaker, and D.J. Ward, editors, *Approximation Theory, VI, pages 203–206*. Academic Press, New York, 1991.

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

[23] R. M. Dudley. *Real analysis and probability*. Mathematics Series. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989.

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

[25] R.M. Dudley. Comments on two preprints: Barron (1991), Jones (1991). Personal communication, March 1991.

[26] N. Dyn. Interpolation of scattered data by radial functions. In C.K. Chui, L.L. Schumaker, and F.I. Utreras, editors, *Topics in multivariate approximation*. Academic Press, New York, 1987.

[27] N. Tishby E. Levin 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.

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

[29] R.L. Eubank. *Spline Smoothing and Nonparametric Regression*, volume 90 of *Statistics, textbooks and monographs*. Marcel Dekker, Basel, 1988.

[30] K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183–192, 1989.

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

[32] F. Girosi. On some extensions of radial basis functions and their applications in artificial intelligence. *Computers Math. Applic.*, 24(12):61–80, 1992.

[33] F. Girosi and G. Anzellotti. Rates of convergence of approximation by translates. A.I. Memo 1288, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[34] 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.

[35] F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. A.I. Memo No. 1430, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[36] 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.

[37] Z. Govindarajulu. *Sequential Statistical Procedures.* Academic Press, 1975.

[38] U. Grenander. On empirical spectral analysis of empirical processes. *Ark. Matemat.*, 1:503–531, 1951.

[39] R.L. Hardy. Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res*, 76:1905–1915, 1971.

[40] R.L. Hardy. Theory and applications of the multiquadric-biharmonic method. *Computers Math. Applic.*, 19(8/9):163–208, 1990.

[41] E. Hartman, K. Keeler, and J.M. Kowalski. Layered neural networks with gaussian hidden units as universal approximators. (submitted for publication), 1989.

[42] 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.

[43] J.A. Hertz, A. Krogh, and R. Palmer. *Introduction to the theory of neural computation.* Addison-Wesley, Redwood City, CA, 1991.

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

[45] 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.

[46] B. Irie and S. Miyake. Capabilities of three-layered Perceptrons. *IEEE International Conference on Neural Networks*, 1:641–648, 1988.

[47] 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 Statistic*, 1990. (to appear).

[48] 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.

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

[50] 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.

[51] A. Lapedes and R. Farber. How neural nets work. In Dana Z. Anderson, editor, *Neural Information Processing Systems*, pages 442–456. Am. Inst. Physics, NY, 1988. Proceedings of the Denver, 1987 Conference.

[52] H. Linhart and W. Zucchini. *Model Selection.* John Wiley and Sons,, 1986.

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

[54] G. G. Lorentz. Metric entropy, widths, and superposition of functions. *Amer. Math. Monthly*, 69:469–485, 1962.

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

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

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

[58] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constr. Approx.*, 2:11–22, 1986.

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

[60] John E. 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, 1992. Morgan Kaufman.

[61] John E. Moody. 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.

[62] P. Niyogi. Active learning of real valued functions. Preprint, 1993.

[63] M. Opper and D. Haussler. Calculation of the learning curve of bayes optimal class algorithm for learning a perceptron with noise. In *Proceedings of COLT, Santa Cruz, CA*, pages 75–87, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[64] A. Pinkus. *N-widths in Approximation Theory.* Springer-Verlag, New York, 1986.

[65] G. Pisier. Remarques sur un resultat non publiè de B. Maurey. In Centre de Mathematique, editor, *Seminarie d'analyse fonctionelle 1980–1981*, Palaiseau, 1981.

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

[67] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

[68] T. Poggio and F. Girosi. Networks for Approximation and Learning. In C. Lau, editor, *Foundations of Neural Networks*, pages 91–106. IEEE Press, Piscataway, NJ, 1992.

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

[70] M.J.D. Powell. The theory of radial basis functions approximation in 1990. Technical Report NA11, Department of Applied Mathematics and Theoretical Physics, Cambridge, England, December 1990.

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

[72] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Ann. Stat.*, 11:416–431, 1983.

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

[74] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(9):533–536, October 1986.

[75] M. Stone. Cross-validatory choice and assessment of statistical predictors(with discussion). *J. R. Statist. Soc.*, B36:111–147, 1974.

[76] R. W. Liu T. P. Chen, H. Chen. A constructive proof of approximation by superposition of sigmoidal functions for neural networks. Preprint, 1990.

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

[78] A.F. Timan. *Theory of approximation of functions of a real variable.* Macmillan, New York, 1963.

[79] L.G. Valiant. A theory of learnable. *Proc. of the 1984 STOC*, pages 436–445, 1984.

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

[81] 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.

[82] V.N. Vapnik and A. Ya. Chervonenkis. The necessary and sufficient conditions for the uniform convergence of averages to their expected values. *Teoriya Veroyatnostei i Ee Primeneniya*, 26(3):543–564, 1981.

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

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

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