Generalization Bounds for Function Approximation from Scattered Noisy Data

Partha Niyogi\textsuperscript{1} and Federico Girosi\textsuperscript{2}

\textsuperscript{1} Artificial Intelligence Laboratory\textsuperscript{*}
545 Technology Square
Massachusetts Institute of Technology
Cambridge, MA 02141, USA
E-mail: niyogi@research.bell-labs.com

\textsuperscript{2} Center for Biological and Computational Learning
45 Carleton Street, E25-201
Massachusetts Institute of Technology
Cambridge, MA, 02141, USA
E-mail: girosi@ai.mit.edu

We consider the problem of approximating functions from scattered data using linear superpositions of non-linearly parameterized functions. We show how the total error (generalization error) can be decomposed into two parts: an approximation part that is due to the finite number of parameters of the approximation scheme used, and an estimation part that is due to the finite number of data available. We bound each of these two parts under certain assumptions and prove a general bound for a class of approximation schemes that include radial basis functions and multilayer perceptrons.

1 Introduction

In this paper we investigate the problem of providing error bounds for approximation of an unknown function from scattered, noisy data. This problem has particular relevance in the field of machine learning, where the unknown function represents the task that has to be learned and the scattered data represents the examples of this task.

An obvious quantity of interest for us is the generalization error – a measure of how much the result of the approximation scheme differs from the unknown function – typically studied as a function of the number of data points. Since the data are randomly generated and noisy, the analysis of the generalization error necessarily involves statistical considerations in addition to the traditional

\textsuperscript{*} Current Address, Bell Labs, Lucent Technologies, Murray Hill, NJ 07974.
approximation theoretic ones. In this paper, we show how the total generalization error can be decomposed into two parts: (1) an approximation component that exists due to the finite dimensionality of the manifold to which the approximating function belongs and (2) an estimation component that is due to the finiteness of the randomly drawn data.

Each of these two components has been investigated in separate contexts in the past. In classical approximation theory, a well known quantity is the degree of approximation, which depends only on the nature of the function being approximated and the class of approximating functions. As the dimensionality of the approximating family (parameterized in some fashion) increases, the approximant converges to the true function. In statistics, one typically assumes the true function belongs to the same family to which the approximating function belongs — but is now known only through a randomly drawn data set. Estimates of this function are constructed from the data and one typically studies the convergence of this estimate to the true function as the data goes to infinity. In this paper, we discuss how approximating from scattered data requires us to deal with both aspects of the problem and provide explicit bounds for a class of approximation techniques. In particular, we will show how one needs to grow the size of the approximating family as a function of the data for convergence to the true target function. We will discuss these issues in greater detail as the paper progresses.

Before proceeding any further, it is worthwhile to introduce some terminology from machine learning that we will use at convenient points in the paper. The function \( f \) that has to be approximated is called the target function. The class \( \mathcal{F} \) to which the target belongs is called the target class. The approximant itself is called the hypothesis function and the class \( \mathcal{H} \) to which it belongs is called the hypothesis class. The learner is exposed to noisy data (examples, or \((x, y)\) pairs) from which it constructs an approximant (hypothesis). From this perspective, the learner is simply an algorithm that maps data to hypotheses, i.e., an approximation scheme. The problem of learning from examples is the problem of approximating the target function from randomly drawn \((x, y)\) pairs using the approximating family \( \mathcal{H} \).

In this paper, we focus on a class of approximation problems where the approximating family \( \mathcal{H} \) can be represented as a linear combination of functions drawn from some fixed finite-dimensional class \( \mathcal{G} \) whose elements can be typically represented by a finite number of parameters. Of course, the class \( \mathcal{H} \) is infinite dimensional and such approximating families include radial basis functions with free centers and multilayer perceptrons — two kinds of “neural network” learning schemes that have received much attention in the machine learning community. We are able to derive generalization error bounds for such schemes and focus in particular on cases where the rate of approximation can be shown to be independent of the number of input variables. We are motivated to look at such cases
due to the fact that machine learning problems (in computer vision for example) typically require a large number of variables where the curse of dimensionality becomes a serious consideration.

The rest of the paper proceeds as follows: in section 2, we describe the approximation problem formally and discuss how the generalization error can be usefully viewed as composed of two parts. We present our main results in section 3, provide formal proofs in section 4 and conclude with some remarks and open problems in section 5.

2 Statement of the Problem

2.1 The General Problem

We discuss various components of the problem of approximating an unknown (target) function from scattered data (examples).

The Target Function

Let \( X \) and \( Y \) be arbitrary subsets of \( \mathbb{R}^k \) and \( \mathbb{R} \) respectively. Let there be a probability distribution \( P \) on the space \( X \times Y \) according to which \((x, y)\) pairs (labeled examples) are drawn in i.i.d. fashion and presented to the learner. In a least squares setting, the learner would ideally like to minimize the “mean squared error” or the expected risk over the class \( \mathcal{F} \). In other words, the learner would like to obtain:

\[
 f_0 = \arg\min_{f \in \mathcal{F}} I[f] = \arg\min_{f \in \mathcal{F}} E[(y - f(x))^2]
\]

Here the expected risk \( I[f] \) is the functional \( E[(y - f(x))^2] \) where the expectation is taken with respect to the distribution \( P \). The solution \( f_0 \), that minimizes the mean squared error (expected risk) is referred to as the regression function. It can easily be shown that

\[
 f_0 = E[y|x] \tag{1}
\]

This \( f_0 \in \mathcal{F} \) represents the target function that the learner wishes to approximate with high accuracy and great confidence. Note that we have implicitly assumed that \( E(y|x) \in \mathcal{F} \).

The Learner’s Estimate

The learner constructs an estimate of the target function \( f_0 \) by searching within a class of parameterized estimators. In general, there are two sources of error for the learner. First, it uses an estimator with a finite number of parameters. Second, it does not have access to the underlying distribution, therefore cannot
compute \( I[f] \). Instead the learner samples the distribution obtaining \( l \) data points and constructs an empirical least squares estimate as follows:

Let \((x_i, y_i); i = 1 \ldots l\) be a data set of \( l \) sample points obtained by sampling \( X \times Y \) in i.i.d. fashion according to the probability distribution \( P \). The empirical risk, \( I_{\text{emp}}[f] \), for any function \( f \) is given by:

\[
I_{\text{emp}}[f] \equiv \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2
\]

Finally, let us denote a generic subset of \( \mathcal{F} \) whose elements are parametrized by a number of parameters proportional to \( n \), by \( H_n \). Moreover, let us assume that the sets \( H_n \) form a nested family, that is

\[
H_1 \subset H_2 \subset \ldots \subset H_n \subset \ldots \subset H.
\]

The estimator is chosen from this class by minimizing the empirical risk over the set \( H_n \). In other words, the learner’s estimate is given by

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

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. In particular, if \( H_n \) is the class of functions which can be represented as \( f = \sum_{\alpha=1}^{n} c_{\alpha} G(x; w_\alpha) \) then eq. (2) can be written as

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

The distance (error) between the target \( f_0 \) and the learner’s estimate \( \hat{f}_{n,l} \) is due to (i) the finite number of parameters \( n \) and (ii) the finite number of data \( l \). This paper bounds this distance under some specific conditions.

2.2 **Approximation and Estimation Errors**

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 it is clear 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 \( I_{\text{emp}} \), obtaining as our final estimate the
function \( \hat{f}_{n,l} \). Our goal is to bound the distance of \( \hat{f}_{n,l} \) (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, the quantity that we need to bound, that we will call generalization error, is:

\[
E[(f_0 - \hat{f}_{n,l})^2] = \int \mathbf{x} P(\mathbf{x})(f_0(\mathbf{x}) - \hat{f}_{n,l}(\mathbf{x}))^2 =
\]

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

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 (11) of proposition 2 as

\[
E[(f_0 - f_n)^2] = I[f_n] - I[f_0].
\]  

(3)

Notice that the approximation error does not depend on the data set \( D_t \), but depends only on the approximating power of the class \( H_n \). 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 \( H = \bigcup_{n=1}^{\infty} H_n \) is dense in \( \mathcal{F} \). In other words, as the number \( n \) of parameters gets larger the representation capacity of \( H_n \) increases, and allows a better and better approximation of the regression function \( f_0 \). The approximation error has been studied extensively in approximation theory, for many different choices of \( H_n \) and \( \mathcal{F} \). More recently, several results have been proved for a class of hypothesis spaces called multilayer perceptron, that seems to be important in practical applications [8, 18, 2, 3, 13, 28, 27].

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* [41]. 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 [42, 43, 41, 44], [36], [12], and [16]. Under a variety of different hypotheses it is possible to prove that, with probability \( 1 - \delta \), a bound of this form is valid:

\[ |I_{\text{emp}}[f] - I[f]| \leq \omega(l, n, \delta) \quad \forall f \in H_n \quad (4) \]

The specific form of \( \omega \) 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 proposition, the generalization error can be bounded by the sum of the two:

**Proposition 1**

The following inequality holds:

\[ \|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)} \leq \varepsilon(n) + 2\omega(l, n, \delta) \quad (5) \]

**Proof:** using the decomposition of the expected risk (11), the generalization error can be written as:

\[ \|f_0 - \hat{f}_{n,l}\|^2_{L^2(P)} = E[(f_0 - \hat{f}_{n,l})^2] = I[\hat{f}_{n,l}] - I[f_0] \quad (6) \]

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

\[ E[(f_0 - \hat{f}_{n,l})^2] \leq |I[\hat{f}_n] - I[f_0]| + |I[f_n] - I[\hat{f}_{n,l}]| \quad (7) \]

In the first term of the right hand side of the previous inequality we recognize the approximation error (3). If a bound of the form (4) is known for the generalization error, it is simple to show that the second term can be bounded as
\[ |I[f_n] - I[\hat{f}_{n,i}]| \leq 2\omega(l, n, \delta) \]

To see this, let us assume that, with probability \(1 - \delta\) a uniform bound has been established:

\[ |I_{\text{emp}}[f] - I[f]| \leq \omega(l, n, \delta) \quad \forall f \in H_n \tag{8} \]

We want to prove that the following inequality also holds:

\[ |I[f_n] - I[\hat{f}_{n,i}]| \leq 2\omega(l, n, \delta) . \tag{9} \]

This is easily established by using the fact that the bound is uniform (so that it holds for both \(f_n\) and \(\hat{f}_{n,i}\)) and by noticing that, by definition of \(f_n\) and \(\hat{f}_{n,i}\) we have:

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

\[ I_{\text{emp}}[\hat{f}_{n,i}] \leq I_{\text{emp}}[f_n] \tag{10} \]

In fact, combining inequalities (8) and (10) we have:

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

However, we also have by (8) that:

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

and combining the last 2 inequalities one obtains:

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

and the result follows.

Thus proposition (1) is proved \(\Box\).

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.