MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1164 October 1989
C.B.I.P. Paper No. 45

Networks and the Best Approximation Property

Federico Girosi and Tomaso Poggio

Abstract

Networks can be considered as approximation schemes. Multilayer networks of the backpropagation type can approximate arbitrarily well continuous functions (Cybenko, 1989; Funahashi, 1989; Stinchcombe and White, 1989). We prove that networks derived from regularization theory and including Radial Basis Functions (Poggio and Girosi, 1989), have a similar property. From the point of view of approximation theory, however, the property of approximating continuous functions arbitrarily well is not sufficient for characterizing good approximation schemes. More critical is the property of *best approximation*. The main result of this paper is that multilayer networks, of the type used in backpropagation, are not best approximation. For regularization networks (in particular Radial Basis Function networks) we prove existence and uniqueness of best approximation.

© Massachusetts Institute of Technology,1994

This paper describes research done within the Center for Biological Information Processing, in the Department of Brain and Cognitive Sciences, and at the Artificial Intelligence Laboratory. This research is sponsored by a grant from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Artificial Intelligence Center of Hughes Aircraft Corporation; by the Alfred P. Sloan Foundation; by the National Science Foundation. Support for the A. I. Laboratory's artificial intelligence research is provided by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76–85–C–0010, and in part by ONR contract N00014–85–K–0124.

1 Introduction

Learning an input-output relation from examples can be considered as the problem of approximating an unknown function $f(x)$ from a set of sparse data points (Poggio and Girosi, 1989). From this point of view, feedforward networks are equivalent to a parametric approximating function $F(W, x)$. As an example, consider a feedforward network, of the multilayer perceptron type, with one hidden layer; the vector W corresponds, then, to the two sets of "weights," from the input to the hidden layer, and from the hidden layer to the output. Even before considering the problem of how to find the appropriate values of W for the set of data, the fundamental representational problem must be approached: which class of mappings f can be approximated by F, and how well? The neural network field has recently seen an increasing awareness of this problem. Several results have been published, all showing that multilayer perceptrons of different form and complexity can approximate arbitrarily well a continuous function, provided that an arbitrarily large number of units is available (Cybenko, 1989; Funahashi, 1989; Moore and Poggio, 1988; Stinchcombe and White, 1989; Carrol and Dickinson, 1989). This property is shared by algebraic and trigonometric polynomials, as is shown by the classical Weierstrass Theorem, and for this reason we shall refer to it as the Weierstrass property. results of this type should not be taken to mean that the approximation scheme is a "good" approximation scheme. An indication of the latter point is provided, in the case of multilayer perceptron networks, of the type used for backpropagation, by a closer look at the published results. Taken together, they imply that almost any nonlinearity at the hidden layer level and a variety of different architectures (one or more hidden layers, for instance) insures the Weierstrass property (Funahashi, 1989; Cybenko, 1989; Stinchcombe and White, 1989). There is nothing special about sigmoids, and in fact many classical approximation schemes exist that can be represented as a network with a hidden layer and that exhibit the Weierstrass property. In a sense this property is not very useful for characterizing approximation schemes, since many schemes have it. Literature in the field of approximation theory reflects this situation, since it emphasizes other properties in characterizing approximation schemes. In particular, a critical concept is that of *best approximation*. An approximation scheme has the best approximation property if in the set A of approximating functions (for instance the set $F(W, x)$ spanned by parameters W) there is one that has minimum distance from any given function of a larger set Φ (a more formal definition is given later). Several questions can be asked, such as the existence, uniqueness, computability, etc., of the best approximation.

In this paper, we show that feedforward multilayer networks of the backpropagation type (Rumelhart et al., 1986, 1986a; Sejnowski and Rosenberg, 1987) do not have the best approximation property for the class of continuous functions defined on a subset of R^n. On the other hand, we prove that for networks derived from regularization, and in particular for radial basis function networks, best approximation exists and is unique. We also prove that these networks approximate arbitrarily well continuous functions (see Appendix B and C). We have recently shown that radial basis function approximation schemes can be derived from regularization and are therefore equivalent to generalized (radial) splines (Poggio and Girosi, 1989). For Radial Basis Function networks we prove existence and uniqueness of best approximation.[1]

The plan of the paper is as follows. We first formalize the previous arguments, then introduce some basic notions from approximation theory. Next, we prove that multilayer networks of the type used for backpropagation do not have the best approximation property, and that networks obtained from regularization theory have this property. In the last section, we discuss the implications of these results and list some open questions. Appendix B proves that the Stone-Weierstrass theorem holds for Gaussian Radial Basis Function networks (with different variances). In appendix C we prove a more general result: regularization networks approximate arbitrarily well any continuous function on a compact subset of R^n.

2 Most networks approximate continuous functions

In recent years there have been attempts to find a mathematical justification for the use of feedforward multilayer networks of the type used for backpropagation. Typical results deal with the possibility, given a network, of approximating any continuous function arbitrarily well. In mathematical terms this means that the set of functions that can be computed by the network is *dense* (see Appendix A) in the space of the continous functions $C[U]$ defined on some subset U of R^d. The most recent results (Cybenko, 1989; Funahashi, 1989; Stinchcombe and White, 1989) consider networks with just one layer of hidden units, that correspond to the following class of approximating functions:

$$\Sigma \equiv \{f \in C[U] \mid f(\mathbf{x}) = \sum_{i=1}^{m} c_i \sigma(\mathbf{x} \cdot \mathbf{w}_i + \theta_i), U \subset R^d, \mathbf{w}_i \in R^d, c_i, \theta_i \in R, m \in N\} \tag{1}$$

where σ is a continuous function. Depending on σ, the set Σ may or may not be dense in the space of the continuous functions. The set $\mathcal{D}$ of functions σ such that Σ is dense seems to be large. For instance, the *sigmoidal* functions, that is functions such that

$$\lim_{t \to +\infty} \sigma(t) = 1$$

$$\lim_{t \to -\infty} \sigma(t) = 0$$

belong to $\mathcal{D}$ (Cybenko, 1989; Funahashi, 1989). Many other types of functions in $\mathcal{D}$ can be found in the paper of Cybenko (1989). The set $\mathcal{D}$ has been recently extended by the result of Stinchcombe and White (1989). In fact they prove that it contains all the functions whose mean value is different from zero and whose L_p-norm is finite for $1 \leq p < \infty$.

Other networks can be built, such that the corresponding set of approximating functions is dense in $C[U]$. Consider for example the network in figure 1. This is the most general network with one layer of hidden units, and the class of approximating functions corresponding to it is

[1] The theory has been extended by introducing the more general schemes of GRBF and HyperBF, which can be considered as the network equivalent of generalized multidimensional splines with free knots.

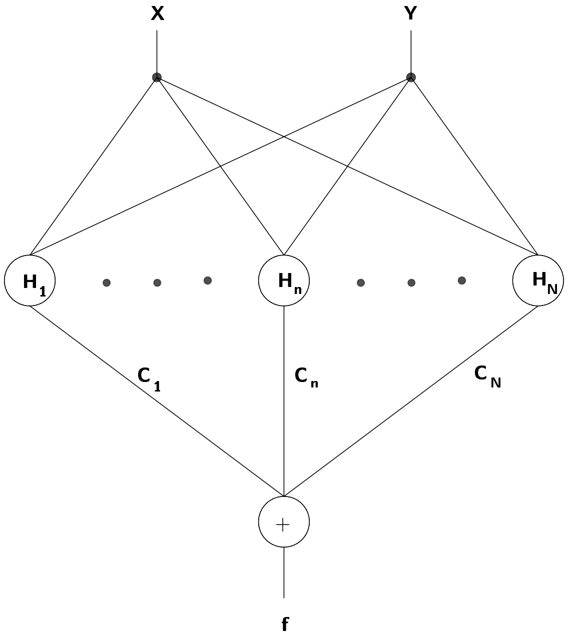

Figure 1: *The most general network with one layer of hidden units. Here we show the two-dimensional case, in which* $\mathbf{x} = (x, y)$. *Each function* H_i *can depend on a set of unknown parameters, that are computed during the learning phase, as well as the coefficients* c_i. *When* $H_i = \sigma(\mathbf{x} \cdot \mathbf{w}_i + \theta_i)$ *a network of the backpropagation type is recovered, while* $H_i = H(\|\mathbf{x} - \mathbf{t}_i\|)$ *corresponds to RBF or GRBF scheme (Broomhead and Lowe, 1988; Poggio and Girosi, 1989).*

$$\mathcal{N} \equiv \{f \in C[U] | f(\mathbf{x}) = \sum_{i=1}^{m} c_i H_i(\mathbf{x}), U \subset R^d, H_i \in C[U], m \in N\}. \tag{2}$$

The function H_i are of the form $H_i = H(\mathbf{x}; \mathbf{W}_i)$, where $\mathbf{W}_i$ is a vector of unknown parameters in some multidimensional space and H is a continuous function. If the H_i are appropriately chosen the set $\mathcal{N}$ can be dense in $C[U]$. For example the H_i could be algebric or trigonometric polynomials, and in this case the denseness of $\mathcal{N}$ would be a trivial consequence of the Stone-Weierstrass theorem (see Appendix B). This theorem allows a significant extension of the set of "basis" functions H_i. Appendix B gives another example, showing how Gaussian functions of radial argument (and different variances) can be used to approximate any continuous function. Appendix C provides a more powerful result showing that *all networks derived from regularization theory can approximate arbitrarily well continuous functions on a compact subset of* R^n. This result includes, in particular, Radial Basis Functions networks with the radial basis function being the Green's function of a self-adjoint differential operator associated to the Tikhonov stabilizer. Such Green's functions include most of the known approximation schemes, such as the Gaussian and several types of splines and many functions, but not all functions, that satisfy some sufficient conditions given by Micchelli (1986) in order to be interpolating functions.

Since a large number of networks can approximate arbitrarily well any continuous functions, it is natural to ask whether this property is really important from the point of view of approximation theory, and whether other more fundamental properties can be characterized. As we mentioned already, one of the basic properties that an approximating set should have is the *best approximation* property, that guarantees that the approximation problem has a solution. The next section focuses our attention on the relationship between this property and different kind of networks, since this seems to be a more appropriate starting point for a complete analysis of the networks performances from a rigorous mathematical point of view.

3 Basic facts in approximation theory

3.1 The best approximation property

An informal formulation of the approximation problem can be stated as follows: *given a function f belonging to some prescribed set of functions Φ, and given a subset A of Φ, find the element a of A that is the "closest" to f.*

In order to give this formulation a precise mathematical meaning, some definitions are needed. First of all a notion of "distance" has to be introduced on the set Φ. Since this set is usually assumed to be a normed linear space, with norm indicated by $\|\cdot\|$, the distance $d(f,g)$ between two elements f and g of Φ is naturally defined as $\|f-g\|$. Given $f \in \Phi$ and $A \subset \Phi$ we can now define the *distance of f from A* as

$$d(f,A) \equiv \inf_{a\in A} \|f-a\|. \tag{3}$$

If the infimum of $\|f-a\|$ is attained for some element a_0 of A, that is if there exists an $a_0 \in A$ such that $\|f-a_0\| = d(f,A)$, this element is said to be a *best approximation to f from A*. A set A is called an *existence set* (*uniqueness set*, resp.) if, to each $f \in \Phi$, there is at least (at most, resp.) one best approximation to f from A. If the set A is an existence set we will also say that it has the best approximation property. A set A is called a *Tchebycheff set* if it is an existence set and a uniqueness set. We are now ready to give a precise formulation of the approximation problem:

Approximation problem: given $f \in \Phi$ and $A \subset \Phi$ find a best approximation to f from A.

From the definition above it is clear that the approximation problem has a solution if and only if A is an existence set, and a large part of approximation theory has been devoted to proving existence theorems, which give sufficient conditions to guarantee existence and possibly uniqueness of closest points. We will only present very simple properties of sets with the best approximation property, and will apply these result to network architectures, in order to understand their properties from the point of view of approximation theory.

We begin with the following observation:

Proposition 3.1 *Every existence set is closed.*

Proof. Let $A \subset \Phi$ be an existence set, and suppose that it is not closed. Then there is a sequence $\{a_n\}$ of elements of A that converges to an element f that is not in A, that is there exists an $f \in \Phi \backslash A$ such that

$$\lim_{n \to \infty} d(f, a_n) = 0$$

This means that $d(f, A) = 0$, and since A is an existence set there is an element $a_0 \in A$ such that $\|f - a_0\| = 0$. By the properties of the norm this implies that $f = a_0$, which is absurd because $f \notin A$ and $a_0 \in A$. Then A must be closed. □

The converse of this proposition is not true, that is closedness is not sufficient for a set to be an existence set. However the stronger condition of *compactness* is sufficient, as the following theorem shows.

Theorem 3.1 *Let A be a compact set in a metric space Φ. Then A is an existence set.*

Proof. For each $f \in \Phi$ the distance $d(f, a)$, with $a \in A$, is a continuous real valued function defined on the compact set A. From theorem A.2 of Appendix A it attains its maximum and minimum value on this set and this concludes the proof. □

In the next section we apply these simple results to some network architectures.

4 Networks and approximation theory

From the point of view of approximation theory a feedforward network is a representation of a set A of parametric functions, and the learning algorithm corresponds to the search of the best approximation to some target function f from A. Since in general a best approximation does not exist unless the set A has some properties (see, for instance, theorem 3.1), it is of interest to understand which classes of networks have these properties.

4.1 Multilayer networks of the backpropagation type do not have the best approximation property

Here we consider the class of networks of the backpropagation type with one layer of hidden units. The space Φ of functions that have to be approximated is chosen to be $C[U]$, the set of continuous functions defined on a subset U of R^d with some unspecified norm. If the number of hidden units is m, the functions that can be computed by such networks belong to the following set σ^m:

$$\sigma^m \equiv \{f \in C[U] \mid f(\mathbf{x}) = \sum_{i=1}^{m} c_i \sigma(\mathbf{x} \cdot \mathbf{w}_i + \theta_i), \mathbf{w}_i \in R^d, c_i, \theta_i \in R\} \quad (4)$$

where $\sigma(x)$ is usually a sigmoidal function. We now show that σ^m *is not* an existence set, and this does not the depend on the norm that has been chosen. The result is proved in the case of σ being a sigmoid and for one hidden layer, $\sigma(x) = (1 + e^{-x})^{-1}$, but *it holds for every other non trivial choice of nonlinear function and for networks with more than one hidden layer.*

Proposition 4.1 *The set σ^m is not an existence set for $m \geq 2$.*

Proof: A necessary condition for a set to be an existence set is to be closed. Therefore it is sufficient to show that σ^m is not closed, and this can be done by showing an accumulation point that does not belong to it. Let us consider the following function:

$$f_\delta(\mathbf{x}) = \frac{1}{\delta}\left(\frac{1}{1+e^{-[\mathbf{w}\cdot\mathbf{x}+\theta]}} - \frac{1}{1+e^{-[\mathbf{w}\cdot\mathbf{x}+(\theta+\delta)]}}\right)$$

Clearly $f_\delta \in \sigma^m, \forall m \geq 2$, but it easily seen that

$$\lim_{\delta\to 0} f_\delta(\mathbf{x}) \equiv g(\mathbf{x}) = \frac{1}{2(1+\cosh[\mathbf{w}\cdot\mathbf{x}+\theta])}$$

and $g \notin \sigma^m, \forall m \geq 2$. For each $m \geq 2$ the function g is then an accumulation point of σ^m but does not belong to it: σ^m can not be closed and this concludes the proof. □

This result reflects a general fact in non linear approximation theory: usually the set of approximating functions is not closed, and its closure must be added to it in order to obtain an existence set. This is the case, for instance, for the approximation by γ-*polynomials* in one dimension, that are replaced by the *extended* γ-*polynomials*, to guarantee the existence of a best approximating element (Braess, 1986; Rice, 1964, 1969; Hobby and Rice, 1967; De Boor, 1969).

4.2 Existence and uniqueness of best approximation for regularization and RBF

One of the possible approaches to the problem of surface reconstruction is given by regularization theory (Tikhonov and Arsenin, 1977; Bertero et al. 1988). Poggio and Girosi (1989) have shown that the solution obtained by means of this method maps into a class of networks with one hidden layer (an instance of which are Radial Basis Function networks or RBF). In fact the solution can always be written in the parametric form:

$$f(\mathbf{x}) = \sum_{i=1}^{m} c_i \phi_i(\mathbf{x}) \tag{5}$$

where the c_i are unknown, m is the number of data points and the ϕ_i are fixed, depending on the nature of the problem and on the data points. More precisely the "basis function" ϕ_i is of the form $\phi_i(\mathbf{x}) = G(\mathbf{x};\mathbf{x}_i)$, where $\mathbf{x}_i$ is a data point and G is the Green's function of some (pseudo)differential operator P (a term belonging to the null space of P can also appear, see Appendix C). In the particular case of radial function $G = G(\|\mathbf{x}-\mathbf{x}_i\|)$ the RBF method is recovered, and the solution of the approximation problem is then a linear superposition of radial Green's functions G "centered" on the data points.

Notice that this function can be computed by a network that is a special case of the one represented in figure 1. The main difference is that in the general case the functions G_i depend on *unknown* parameters, while in the regularization context only the coefficient c_i are unknown.

Equation 5 means that the approximated solution belongs to the subset T^m of $C[U]$:

$$T^m \equiv \{f \in C[U] \mid f(\mathbf{x}) = \sum_{i=1}^{m} c_i \phi_i(\mathbf{x}), c_i \in R\} \tag{6}$$

Since we have shown that the set of approximating functions associated with networks with one hidden layer of the type used for backpropagation does not have the best approximation property, it is natural to ask whether or not the set T^m has this property [2]. The answer is positive, as is stated in the following proposition:

Proposition 4.2 *The set T^m is an existence set for $m \geq 1$*

Proof. Let f be a prescribed element of $C[U]$, and let a_0 be an arbitrary point of T^m. We are looking for the closest point to f in T^m. It has to lie in the set

$$\{a \in T^m \mid \|a - f\| \leq \|a_0 - f\|\}.$$

This set is clearly closed and bounded, and by theorem A.1 it is compact. The best approximation property comes from theorem 3.1. □

From this proposition we can see that every time that the approximating function is a finite linear combination of basis functions, the set that is spanned by these basis functions is an existence set for $C[U]$. Depending on the norm that is chosen in $C[U]$ the best approximating element can be unique. In fact the following theorem holds (see Appendix A for the definition of *strictly convex*):

Proposition 4.3 *The set T^m, $m \geq 1$ is a Tchebycheff set if the normed space $C[U]$ is strictly convex.*

Proof. The existence has already been proved. Suppose then that there are two best approximating elements f and f' from T^m to a function $g \in C[U]$. Let λ be the distance of g from T^m. Applying the triangular inequality we obtain :

$$\|\frac{1}{2}(f + f') - g\| \leq \frac{1}{2}\|f - g\| + \frac{1}{2}\|f' - g\| = \lambda \tag{7}$$

Since T^m is a vector space, then $\frac{1}{2}(f + f') \in T^m$ and by definition of λ it follows that $\|\frac{1}{2}(f + f')\| \geq \lambda$. This implies that the equality holds in equation 7. If $\lambda = 0$ it is clear that $f = f' = g$. If $\lambda \neq 0$, then we can write equation 7 as

$$\|\frac{1}{2}\left[\frac{(f - g)}{\lambda} + \frac{(f' - g)}{\lambda}\right]\| = 1. \tag{8}$$

This means that the vectors $\frac{(f-g)}{\lambda}$, $\frac{(f'-g)}{\lambda}$ and their midpoints are all of norm 1, but since strcty convexity holds, then $f = f'$. □

Since it is well known that $C[U]$ with the L_p-norms, $1 < p < \infty$ is strictly convex (Rice, 1964), we have then shown that in most cases regularization theory gives an approximating set with the best approximation property and with a unique best approximating element.

[2] Notice that multilayer perceptrons of the type used for backpropagation cannot be derived from any regularization scheme since it cannot be written as the linear superposition of Green's functions of any kind.

5 Conclusions

5.1 GRBF and Best Approximation

We have recently extended the scheme of equation 5 to the case in which the number of basis functions is less than the number of data points (Poggio and Girosi, 1989; Broomhead and Lowe, 1988). The reason for this is that when the number of data points becomes large the complexity of the network may become too high, being proportional to the number of data points. A solution to the approximation problem is sought of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i G(\mathbf{x}; \mathbf{t}_i) \tag{9}$$

where n is smaller than the number of data points and the positions of the "centers" $\mathbf{t}_i$ of the expansion are unknown, having to be found during the learning stage. Does the best approximation property hold for this approximation scheme, that we call Generalized Radial Basis Function (GBRF) method? The answer is no, exactly as for splines with free knots, to which equation 9 is in fact equivalent. By the same arguments we have used in section 4.1 we could show that the set G^n of approximating functions generated by equation 9 (the analogous of the set T^m) is not closed. The scheme, however, has almost the best approximation property in the following sense. The scheme already works satisfactorily if the centers $\mathbf{t}_i$ are fixed to a subset of examples or other positions. In this case G^n is a linear space, and it is an existence set, as well as T^m. We could then have an algorithm in which first the centers are found independently (for instance by the K-means algorithm, see Moody and Darken, 1989) and then the c_i are obtained with gradient descent methods (see Poggio and Girosi, 1989). In this scheme the best approximation property is preserved, while the computational complexity has been reduced with respect to the exact solution of the regularization problem.

There are other ways to make GRBF a best approximation. The most interesting approach is to follow the theory of γ-polynomials (Braess, 1986; Rice, 1964, 1969; Hobby and Rice, 1967; De Boor, 1969) and complete the sets of basis functions with its closure, consisting of an appropriate number of derivatives of the Green's function with respect to its parameters, yielding a best approximation scheme. It seems very difficult to use either of these two approaches for networks of the type used for backpropagation.

5.2 Open Questions

We have not explored the practical consequences of the fact that multilayer networks of the backpropagation type are not best approximation. Intuitively, it seems that the lack of the best approximation property is related to possible practical degeneracies of the solution. In certain situations, because of the fact that the sigmoid, which is asymptotically constant, contains as an argument one set of parameters (the w_i), the precise values of these parameters may not have any significant effect on the output of the network. The same situation happens for GRBF when the centers inside the Green's function are unknown. In the GRBF case, however, we can freeze the $\mathbf{t}_i$ to reasonable values whereas this is impossible in the backpropagation case.

Other questions remain open as well. The most important questions from the viewpoint of approximation theory are: (1) the computation of the best approximation, i.e., which algorithm to use, (2) *a priori* bounds on the goodness of the approximation given some generic information on the class of functions to be approximated, and (3) *a priori* estimates of the complexity of the best approximation, again given generic information on the class of functions to be approximated. In the case of RBF, the latter question is directly related to the size of the required training set, and therefore to the deep issue of sample complexity (see Poggio and Girosi, 1989, section 9.3). About problems 1) and 2) notice that in practical cases it may be admissible to use a scheme which is not best approximation, *if* it provides an almost as good approximation at a much lower computational cost.

Acknowledgments *We are grateful to G. Palm and E. Grimson for useful suggestions. This report describes research done within the Artificial Intelligence Laboratory and the Center for Biological Information Processing in the Department of Brain and Cognitive Sciences. Support for this research is provided by a grant from ONR, Cognitive and Neural Sciences Division, by the Artificial Intelligence Center of Hughes Aircraft Corporation and by the NATO Scientific Affairs Division (0403/87). Support for the A.I. Laboratory's artificial intelligence research is provided by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76–85–C–0010 and in part under Office of Naval Research (ONR) contract N00014–85–K–0124. TP is supported by the Uncas and Ellen Whitaker chair.*

A Definitions and basic theorems

We review here some of the definitions that have been used in the paper. Every set will be assumed to have the structure of metric space, unless differently specified, and the concepts of limit point, infimum and supremum are assumed to be known. All these definitions and theorems can be found in any standard text on functional analysis (Yosida, 1974; Rudin, 1973) and in many books on approximation theory (Braess, 1986; Cheney, 1981).

An important concept is that of *closure*:

Definition A.1 *If $\mathcal{S}$ is a set of elements, then by the closure $[\mathcal{S}]$ of $\mathcal{S}$ we mean the set of all points in $\mathcal{S}$ together with the set of all limit points of $\mathcal{S}$.*

We can now define the *closed* sets as following:

Definition A.2 *A set $\mathcal{S}$ is closed if it is coincident with its closure $[\mathcal{S}]$.*

A closed set then contains all its limit points. Another important definition related to the concept of closure is that of *dense* sets:

Definition A.3 *Let $\mathcal{T}$ a subset of the set $\mathcal{S}$. $\mathcal{T}$ is dense in $\mathcal{S}$ if $[\mathcal{T}] = \mathcal{S}$.*

If $\mathcal{T}$ is dense in $\mathcal{S}$ then each element of $\mathcal{S}$ can be approximated arbitrarily well by elements of $\mathcal{T}$. As an example we mention the set of rational numbers, that is dense in the set of real numbers, and the set of polynomials that is dense in the space of continuous functions (see appendix B).

In order to extend some properties of the real valued functions defined on an interval to real valued functions defined on more complex metric spaces it is fundamental to define the *compact* sets:

Definition A.4 *A compact set is one in which every infinite subset contains at least one limit point.*

It can be shown that, in finite dimensional metric spaces, there exists a simple characterization of compacts sets. In fact the following theorem holds:

Theorem A.1 *Every closed, bounded, finite-dimensional set in a metric linear space is compact.*

The well known Weierstrass theorem on the attainment of the extrema of a continuous function on an interval can now be extended as following:

Theorem A.2 *A continuous real valued function defined on a compact set in a metric space achieves its infimum and supremum on that set.*

A subset of the metric spaces is given by the normed spaces, and among the normed spaces, a special role is played by the *strictly convex* spaces:

Definition A.5 *A normed space is strictly convex if:*

$$\|f\| = \|g\| = \|\frac{1}{2}(f+g)\| = 1 \Rightarrow f = g$$

The geometrical interpretation of this definition is that a space is strictly convex if the unit sphere does not contain any line segment on its surface.

B Gaussian networks and Stone's theorem

It has been proved (Cybenko, 1989; Funahashi, 1989) that a network with a one hidden layer of sigmoidal units can approximate a continuous function arbitrarily well. Here we show that this property, which is well known for algebraic and trigonometric polynomial approximation schemes, is shared by a network with Gaussian hidden units. The proof is a simple application of the Stone-Weierstrass theorem, which is the generalization given by Stone of the Weierstrass approximation theorem (Stone, 1937, 1948). Our result was obtained independently from the equivalent proof of Hartman, Keeler and Kowalski (1989). We first need the definitions of *algebra*.

Definition B.1 *An algebra is a set of elements denoted by $\mathcal{Y}$, together with a scalar field $\mathcal{F}$, which is closed under the binary operators of $+$ (addition between elements of $\mathcal{Y}$), $\times$ (multiplication of elements of $\mathcal{Y}$), $\cdot$ (multiplication of elements in $\mathcal{Y}$ by elements from the scalar field $\mathcal{F}$), such that*

1. *$\mathcal{Y}$ together with $\mathcal{F}$, $+$ and $\cdot$ forms a linear space,*

2. *if $f,\ g,\ h$ are in $\mathcal{Y}$, α is in $\mathcal{F}$, then*

   a. *$f \times g$ is in $\mathcal{Y}$,*
   b. *$f \times (g \times h) = (f \times g) \times h$,*
   c. *$f \times (g + h) = f \times g + f \times h$,*
   d. *$(f + g) \times h = f \times h + g \times h$,*
   e. *$\alpha(f \times g) = (\alpha f) \times g = f \times (\alpha g)$.*

It is an elementary calculation to show that if U is some subsect of R^d then $C[U]$ is an algebra with respect to the scalar field R. We can now define a *subalgebra* as following:

Definition B.2 *A set $\mathcal{S}$ is a subalgebra of the algebra $\mathcal{Y}$ if*

1. *$\mathcal{S}$ is a linear subspace of $\mathcal{Y}$,*

2. *$\mathcal{S}$ is closed under the operation $\times$. That is, if f and g are in $\mathcal{S}$, then $f \times g$ is also in $\mathcal{S}$.*

We can now formulate the Stone's theorem:

Theorem B.1 (Stone, 1937) *Let X be a compact metric space, $C[X]$ the set of continuous functions defined on X, and A a subalgebra of $C[X]$ with the following two properties:*

1. *the function $f(x) = 1$ belongs to A;*

2. *for any two distinct points x and y in X there is a function $f \in A$ such that $f(x) \neq f(y)$.*

Then A is dense in $C[X]$.

As a simple application of this theorem we consider the set of gaussian superpositions, defined as

$$\mathcal{G}_{\mathcal{X}} \equiv \{f \in C[X] \mid f(\mathbf{x}) = \sum_{i=1}^{m} c_i e^{-\frac{(\mathbf{x}-\mathbf{t}_i)^2}{\sigma_i^2}}, X \subset R^d, \mathbf{t}_i \in R^d, c_i, \sigma_i \in R, m \in N\} \tag{10}$$

We can now enunciate the following:

Proposition B.1 *The set $\mathcal{G}_{\mathcal{X}}$ is dense in $C[X]$, where X is a compact subset of R^d.*

Proof: In order to use Stone's therorem, we first have to show that $\mathcal{G}_\mathcal{X}$ is a subalgebra of $C[X]$, for each compact subset X of R^d. The set $\mathcal{G}_\mathcal{X}$ will be a subalgebra of $C[X]$ if the product of two of its elements yields another element of $\mathcal{G}_\mathcal{X}$. Since $\mathcal{G}_\mathcal{X}$ is a linear superposition of gaussians of different variance and centered on different points it is sufficient to deal with the product of two gaussians. From the identity below it follows that the product of two gaussians centered on two points $\mathbf{t}_1$ and $\mathbf{t}_2$ is proportional to a Gaussian centered on a point $\mathbf{t}_3$ that is a convex linear combination of $\mathbf{t}_1$ and $\mathbf{t}_2$. In fact we have:

$$e^{-\frac{(\mathbf{x}-\mathbf{t}_1)^2}{\sigma_1^2}} \cdot e^{-\frac{(\mathbf{x}-\mathbf{t}_2)^2}{\sigma_2^2}} = ce^{-\frac{(\mathbf{x}-\mathbf{t}_3)^2}{\sigma_3^2}} ,$$

$$\mathbf{t}_3 = \frac{\sigma_2^2\mathbf{t}_1 + \sigma_1^2\mathbf{t}_2}{\sigma_1^2 + \sigma_2^2}, \quad \sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad c = e^{-\frac{(\mathbf{t}_1-\mathbf{t}_2)^2}{\sigma_3^2}} .$$

The function $f(\mathbf{x}) = 1$ belongs to $\mathcal{G}_\mathcal{X}$, since it can be considered as gaussian of infinite variance, and for any distinct points $\mathbf{x}, \mathbf{y}$ we can obviously find a function in $\mathcal{G}_\mathcal{X}$ such that $f(\mathbf{x}) \neq f(\mathbf{y})$: the conditions of Stone's theorem are then satisfied and $\mathcal{G}_\mathcal{X}$ is dense in $C[X]$ $\Box$.

C Regularization networks can approximate smooth functions arbitrarily well

In this appendix we briefly describe the regularization method for approximating functions and show that the networks that are derived from a regularization principle can approximate arbitrarily well continuous functions defined on a compact subset of R^n.

Let $S = \{(\mathbf{x}_i, y_i) \in R^n \times R | i = 1, ...N\}$ be a set of data that we want to approximate by means of a function f. The regularization approach (Tikhonov, 1963; Tikhonov and Arsenin, 1977; Morozov, 1984; Bertero, 1986) consists in computing the function f that minimizes the functional

$$H[f] = \sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2 + \lambda\|Pf\|^2$$

where P is a constraint operator (usually a differential operator), $\|\cdot\|^2$ is a norm on the function space to whom Pf belongs (usually the L^2 norm) and λ is a positive real number, the so called *regularization parameter*. The structure of the operator P embodies the a priori knowledge about the solution, and therefore depends on the nature of the particular problem that has to be solved. The general form of the solution of this variational problem is given by the following expansion (Poggio and Girosi, 1989):

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i G(\mathbf{x}; \mathbf{x}_i) + p(\mathbf{x}) \tag{11}$$

where G is the Green's function of the differential operator $\hat{P}P$, $\hat{P}$ being the adjoint operator of P, $p(\mathbf{x})$ is a linear combination of functions that span the null space of P, and the coefficients c_i can be found by inverting a matrix that depends on the data points (Poggio and Girosi, 1989). We remind the reader that the Green's function of an operator $\hat{P}P$ is the function that satisfies the following differential equation (in the distributions sense):

$$\hat{P}P\ G(\mathbf{x};\mathbf{y}) = \delta(\mathbf{x}-\mathbf{y})\ . \tag{12}$$

It is clear that there is a correspondence between the class of functions that can be written in the form (11) (for any number of data points and for any Green's functions G of a self-adjoint operator) and a subclass of feedforward networks with one layer of hidden units, of the type shown in figure 1. Under mild assumptions on $\hat{P}P$, these networks can approximate continuous functions arbitrarily well, as is stated in the following proposition:

Proposition C.1 *For every continuous function F defined on a compact subset of R^n and every piecewise continuous G which is the Green's function of a self-adjoint differential operator, there exists a function $f^*(\mathbf{x}) = \sum_{i=1}^N c_i G(\mathbf{x};\mathbf{x}_i)$, such that for all $\mathbf{x}$ and any positive ϵ the following inequality holds:*

$$|F(\mathbf{x}) - f^*(\mathbf{x})| < \epsilon$$

Proof: Let F be a continuous function defined on a compact set $D \subset R^n$. Its domain of definition can be extended to all R^n by assigning zero value to all points that do not belong to D. The resulting function, that we still call F, is a continous function with bounded support[3]. Consider the space K of "test functions" (Gelfand and Shilov, 1964), that consists of real functions $\phi(\mathbf{x})$ with continuous derivatives of all orders and with bounded support (which means that the function and all its derivatives vanish outside of some bounded region). As Gelfand and Shilov show (Appendix 1.1), there always exists a function $\phi(\mathbf{x})$ in K arbitrarily close to F, i.e. , such that for all $\mathbf{x}$ and for any $\epsilon > 0$,

$$|F(\mathbf{x}) - \phi(\mathbf{x})| < \epsilon.$$

Thus it is sufficient to show that every function $\phi(\mathbf{x}) \in K$ can be approximated arbitrarily well by a linear superposition of Green's functions (function f^* of proposition C.1).

We start with the identity

$$\phi(\mathbf{x}) = \int d\mathbf{y}\phi(\mathbf{y})\delta(\mathbf{x}-\mathbf{y}) \tag{13}$$

where the integral is actually taken only over the bounded region in which $\phi(\mathbf{x})$ fails to vanish. By means of equation 12 we obtain

$$\phi(\mathbf{x}) = \int d\mathbf{y}\phi(\mathbf{y})(\hat{P}PG)(\mathbf{x};\mathbf{y}) \tag{14}$$

and since $\phi(\mathbf{x})$ is in K and $\hat{P}P$ is formally self-adjoint we have

$$\phi(\mathbf{x}) = \int d\mathbf{y}G(\mathbf{x};\mathbf{y})(\hat{P}P\phi)(\mathbf{y}). \tag{15}$$

[3] The support of a continuous function $F(\mathbf{x})$ is the closure of the set on which $F(\mathbf{x}) \neq 0$.

We can rewrite equation 15 as

$$\phi(\mathbf{x}) = \int d\mathbf{y} G(\mathbf{x};\mathbf{y})\psi(\mathbf{y}) \tag{16}$$

where $\psi(\mathbf{x}) = \hat{P}P\phi(\mathbf{x})$. Since $G(\mathbf{x};\mathbf{y})\psi(\mathbf{y})$ is piecewise continuous on a closed domain, this integral exists in the sense of Riemann. By definition of Riemann integral, equation 16 can then be written as

$$\phi(\mathbf{x}) = \Delta^n \sum_{k\in I} \psi(\mathbf{x}_k)G(\mathbf{x};\mathbf{x}_k) + E_{\mathbf{x}}(\Delta) \tag{17}$$

where $\mathbf{x}_k$ are points of a square grid of spacing Δ, I is the finite set of lattice points where $\psi(\mathbf{x}) \neq 0$, and $E_{\mathbf{x}}(\Delta)$ is the discretization error, with the property

$$\lim_{\Delta\to 0} E_{\mathbf{x}}(\Delta) = 0. \tag{18}$$

If we now choose $f^*(\mathbf{x}) = \Delta^n \sum_{k\in I} \psi(\mathbf{x}_k)G(\mathbf{x};\mathbf{x}_k)$, combining equation 18 and equation 17 we obtain

$$\lim_{\Delta\to 0}[\phi(\mathbf{x}) - f^*(\mathbf{x})] = 0. \tag{19}$$

Thus every function $\phi \in K$ can be approximated arbitrarily well by a linear superposition of Green's functions G of a self-adjoint operator, and this concludes the proof □.

Remark: The conditions of proposition C.1 exclude Green's functions that have singularities in the origin. An example is the Green's function associated with the "membrane" stabilizer $P = \vec{\nabla}$ in 2 or more dimensions. In 2 dimensions, the membrane Green's function is $G(r) = -logr$, where $r = \|\mathbf{x} - \mathbf{x}_i\|$ (in 1 dimension $G(x) = |x|$, satisfies the conditions of proposition C.1).

Remark: Notice that in order to approximate arbitrarily well any continous function on a compact domain with functions of the type 11, it is not necessary to include the term p belonging to the null space of P.

References

[1] M. Bertero. Regularization methods for linear inverse problems. In C. G. Talenti, editor, *Inverse Problems*. Springer-Verlag, Berlin, 1986.

[2] M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76:869–889, 1988.

[3] D. Braess. *Nonlinear Approximation Theory*. Springer-Verlag, Berlin, 1986.

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

[5] S.M. Carrol and B.W. Dickinson. Construction of neural nets using the Radon transform. In *Proceedings of the International Joint Conference on Neural Networks*, pages I–607–I–611, Washington D.C., June 1989. IEEE TAB Neural Network Committee.

[6] E.W. Cheney. *Introduction to approximation theory.* Chelsea Publishing Company, New York, 1981.

[7] G. Cybenko. Continuous valued neural networks with two hidden layers are sufficient. Technical report, Dept. of Computer Sciences, Tufts Univ., Medford, MA, 1988.

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

[9] C. de Boor. On the approximation by γ-Polynomials. In I.J. Schoenberg, editor, *Approximation with special emphasis on spline functions*, pages 157–183. Academic Press, New York, 1969.

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

[11] I.M. Gelfand and G.E. Shilov. *Generalized functions. Vol. 1: Properties and Operations.* Academic Press, New York, 1964.

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

[13] C.R. Hobby and J.R. Rice. Approximation from a curve of functions. *Arch. Rat. Mech. Anal.*, 27:91–106, 1967.

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

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

[16] B. Moore and T. Poggio. Representations properties of multilayer feedforward networks. In *Abstracts of the first annual INNS meeting*, page 502, New York, 1988. Pergamon Press.

[17] V.A. Morozov. *Methods for solving incorrectly posed problems.* Springer-Verlag, Berlin, 1984.

[18] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[19] J.R. Rice. *The approximation of functions, Vol. 1.* Addison-Wesley, Reading, MA, 1964.

[20] J.R. Rice. *The approximation of functions, Vol. 2.* Addison-Wesley, Reading, MA, 1969.

[21] W. Rudin. *Functional Analysis.* McGraw-Hill, New York, 1973.

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing*, chapter 8, pages 318–362. MIT Press, Cambridge, MA, 1986.

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

[24] T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce english text. *Complex Systems*, 1:145–168, 1987.

[25] M. Stinchcombe and H. White. Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In *Proceedings of the International Joint Conference on Neural Networks*, pages I–607–I–611, Washington D.C., June 1989. IEEE TAB Neural Network Committee.

[26] M.H. Stone. Applications of the theory of Boolean rings to general topology. *AMS Transactions*, 41:375–481, 1937.

[27] M.H. Stone. The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21:167–183, 237–254, 1948.

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

[29] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems.* W. H. Winston, Washington, D.C., 1977.

[30] K. Yosida. *Functional Analysis.* Springer, Berlin, 1974.