Approximation of Multivariate Functions

Vladimir Ya. Lin and Allan Pinkus

Abstract. We discuss one approach to the problem of approximating functions of many variables which is truly multivariate in character. This approach is based on superpositions of functions with infinite families of smooth simple functions.

§1. Introduction and Motivation

There are numerous methods of approximating functions of many variables. For example, we have the more classic methods using Polynomials, Fourier Series, or Tensor Products, and more modern methods using Wavelets, Radial Basis Functions, Multivariate Splines, or Ridge Functions. Many of these are natural generalizations of methods developed for approximating univariate functions. However functions of many variables are fundamentally different from functions of one variable, and approximation techniques for such problems are much less developed and understood. We will discuss in these few pages one approach to this problem which is truly multivariate in character.

Hilbert's 13th problem, although not formulated in the following terms, was interpreted by some as conjecturing that not all functions of 3 variables could be represented as *superpositions* (compositions) and *sums* of functions of two varaibles. Surprisingly it turned that all functions could be so represented, and even more was true. Kolmogorov and his student Arnold proved in a series of papers in the late 50's that there exist fixed continuous one variable functions h_{ij} such that every continuous function f of n variables on $[0,1]^n$ could be represented in the form

$$f(x_1,\dots,x_n)=\sum_{i=1}^{2n+1} g_i\Big(\sum_{j=1}^{n} h_{ij}(x_j)\Big)$$

where one chooses the continuous one variable functions g_i.

As with any such surprising and deep result, numerous questions and results were spawned. One question asked had to do with smoothness properties of the h_{ij}. Assuming f is in some class of smooth functions, can one choose smooth h_{ij}?

Different answers in different frameworks were given. One answer due to Vitushkin and Henkin (separately and together) was given in the mid 60's (see [5] for a survey of their results). Since the answer is in the negative, we only formulate it for functions of two variables x and y in $[0,1]^2$.

Theorem A. *For any m fixed continuous functions $\psi_i(x,y)$, $i=1,\ldots,m$, and continuously differentiable functions $\phi_i(x,y)$, $i=1,\ldots,m$, the set of functions*

$$\Big\{ \sum_{i=1}^{m} \psi_i(x,y)\, g_i\big(\phi_i(x,y)\big) \,:\, g_i \text{ continuous} \Big\}$$

is nowhere dense in the space of all functions continuous in $[0,1]^2$ with the topology of uniform convergence.

"If" such a theorem were not true, *i.e.*, if such a representation did in fact exist with smooth ϕ_i, then multivariate approximation theory might well look different today. If the ψ_i and ϕ_i were calculable (once and for all), then we could reduce many multivariable problems to 1-variable problems. However, this is all idle speculation as in the original Kolmogorov-Arnold result, the h_{ij} are not C^1 and not at all calculable.

The idea of using superpositions (or compositions) of functions with smooth, nice, functions is a good idea. Let us try to formulate somewhat more precisely what we have in mind.

To make things more concrete, we will deal with the space of continuous real-valued functions defined on $\mathbb{R}^n$, endowed with the topology of uniform convergence on compact sets. (Note that we are not dealing with the topology of uniform convergence on all of $\mathbb{R}^n$.)

Let Φ be a family of continuous (smooth) functions $\phi : \mathbb{R}^n \to \mathbb{R}$. Let us consider the space

$$\text{span}\{\, g(\phi(\mathbf{x})) \,:\, g \in C(\mathbb{R})\,, \phi \in \Phi\}\,,$$

where $\mathbf{x} = (x_1,\ldots,x_n)$ and the g are functions of one variable. Since representation is, as we have seen, seldom possible using a finite number of smooth interior functions, we will here deal with the related (but different) problem of approximation.

The question we will here discuss is that of density. That is when, for any given $f \in C(\mathbb{R}^n)$, any compact set $K \subset \mathbb{R}^n$, and $\varepsilon > 0$, we can find $g_i \in C(\mathbb{R})$ and $\phi_i \in \Phi$, $i=1,\ldots,m$ (m finite but free to be chosen) such that

$$\max_{\mathbf{x}\in K} \Big| f(\mathbf{x}) - \sum_{i=1}^{m} g_i(\phi_i(\mathbf{x})) \Big| < \varepsilon\,.$$

We will look at two general classes of functions Φ, and at the related question of when we can also restrict the set of permissible g.

§2. Translation

Let φ be any fixed C^1 function defined on $\mathbb{R}^n$. Consider

$$\mathcal{M}_\varphi = \operatorname{span}\{g\big(\varphi(\mathbf{x}-\mathbf{a})\big) \,:\, g \in C(\mathbb{R}),\, \mathbf{a} \in \mathbb{R}^n\}\,.$$

That is, we fix one and only one function φ, and let

$$\Phi = \{\varphi(\cdot - \mathbf{a})\}$$

be the set of all translates of this function.

What can we say in this generality?

a) If φ is such that there exists $c < d$ such that

$$\{\mathbf{x} \,:\, c \le \varphi(\mathbf{x}) \le d\}$$

is a bounded nonempty set, then $\overline{\mathcal{M}_\varphi} = C(\mathbb{R}^n)$. (In this case, we can also fix a particular g.) Thus we have density if, for example, $\varphi(\mathbf{x}) = \|\mathbf{x}\|$ and the norm $\|\cdot\|$ is smooth (or some power of the norm is smooth). This condition is *not* necessary.

b) A necessary condition for density is that

$$\operatorname{span}\{\varphi(\mathbf{x}-\mathbf{a}) : \mathbf{a} \in \mathbb{R}^n\}$$

must separate points. In other words, for $\mathbf{x} \neq \mathbf{y}$ there must exist points $\mathbf{a}$ and $\mathbf{b}$ such that $\varphi(\mathbf{x}-\mathbf{a}) \neq \varphi(\mathbf{y}-\mathbf{b})$. If φ is a polynomial then we do not have the separation of points if φ is a polynomial of less than n variables. Aside from a linear change of variables, this is the only case where the above set does not separate points. That is, if φ is an algrebraic polynomial

$$\operatorname{span}\{\varphi(\mathbf{x}-\mathbf{a}) : \mathbf{a} \in \mathbb{R}^n\}$$

separates points *if and only if* there does not exist a non-singular $n \times n$ matrix C such that φ is independent of $(C\mathbf{x})_n$. Based on various cases which have been studied in some detail, we would like to conjecture that if φ is a polynomial and

$$\operatorname{span}\{\varphi(\mathbf{x}-\mathbf{a}) : \mathbf{a} \in \mathbb{R}^n\}$$

separates points, then $\overline{\mathcal{M}_\varphi} = C(\mathbb{R}^n)$.

Two results permit us to easily restrict the size of the set of the approximants without losing the density property. The first is specific, the second general.

c) If φ is a polynomial, then it is not necessary to consider *all* translates. In fact, if $\mathcal{A}$ is any subset of $\mathbb{R}^n$ for which no non-trivial polynomial vanishes on $\mathcal{A}$, then

$$\overline{\operatorname{span}}\{g\big(\varphi(\mathbf{x}-\mathbf{a})\big) : g \in C(\mathbb{R}),\, \mathbf{a} \in \mathbb{R}^n\} = \overline{\operatorname{span}}\{g\big(\varphi(\mathbf{x}-\mathbf{a})\big) : g \in C(\mathbb{R}),\, \mathbf{a} \in \mathcal{A}\}\,.$$

Thus, in case φ is a polynomial, we can, for example, consider shifts only by elements of $\mathbb{Z}^n$, or alternatively by all elements of any set with interior.

d) We can always replace *all* $g \in C(\mathbb{R})$ in the above by *all* $g \in \mathcal{B}$ where $\mathcal{B}$ is any dense subset of $C(\mathbb{R})$ (in the requisite topology). For example, we can let g run over the set of all monomials. A more interesting example is the following:

Let $\sigma \in C(\mathbb{R})$. In the topology of uniform convergence on compact subsets of $\mathbb{R}$, it may be shown that

$$C(\mathbb{R}) = \overline{\text{span}}\{\sigma(at+b) \,:\, a, b \in \mathbb{R}\}$$

if and only if σ is *not* a polynomial (see Leshno, Lin, Pinkus and Schocken [2]).

It follows quite easily from this result that for any set Φ, and any $\sigma \in C(\mathbb{R})$ which is not a polynomial,

$$\overline{\text{span}}\{g\big(\phi(\mathbf{x})\big) : g \in C(\mathbb{R}),\, \phi \in \Phi\} = \overline{\text{span}}\{\sigma\big(a\phi(\mathbf{x}) + b\big) \,:\, a, b \in \mathbb{R},\, \phi \in \Phi\}\,.$$

Sometimes it is possible to further restrict this set in that we need not run over all $a, b \in \mathbb{R}$. For example, if σ is what is called a *sigmoidal function*. That is, $\sigma \in C(\mathbb{R})$ and

$$\lim_{x \to -\infty} \sigma(x) = 0 \text{ and } \lim_{x \to \infty} \sigma(x) = 1,$$

then we can replace $a, b \in \mathbb{R}$ in the above by $k, \ell \in \mathbb{Z}$ (see Chui, Li [1]). It also always suffices to restrict a to any non-trivial interval.

Thus, for example, if σ is *sigmoidal* and $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$, then

$$C(\mathbb{R}^n) = \overline{\text{span}}\{\sigma(k\|\mathbf{x} - \mathbf{m}\| + \ell) \,:\, k, \ell \in \mathbb{Z},\, \mathbf{m} \in \mathbb{Z}^n\}\,.$$

e) One particular choice of φ is of special interest. Consider $n = 2$, and $\varphi(x, y) = x^2 + y^2$. In other words, we are considering the space

$$\mathcal{M}_\varphi = \text{span}\{g\big((x-a)^2 + (y-b)^2\big) \,:\, g \in C(\mathbb{R}),\, (a, b) \in \mathbb{R}^2\}\,.$$

(Such functions are *Radial Functions*.) Instead of considering translates (centering) by all $(a, b) \in \mathbb{R}^2$, let us consider translates only by $(a, b) \in \mathcal{A}$, where $\mathcal{A}$ is some subset of $\mathbb{R}^2$.
From (c) we know that $\overline{\mathcal{M}_\varphi} = C(\mathbb{R}^2)$ if no non-trivial polynomial vanishes on $\mathcal{A}$. However, this is far from necessary. Let us consider some simple sets $\mathcal{A}$ which *are* zero sets of non-trivial polynomials.

(i) If $\mathcal{A}$ is only a finite number of points, then by Theorem A, $\overline{\mathcal{M}_\varphi} \neq C(\mathbb{R}^2)$.

(ii) If $\mathcal{A}$ is only a *straight line* (or any subinterval thereof), then $\overline{\mathcal{M}_\varphi} \neq C(\mathbb{R}^2)$. However, $\mathcal{M}_\varphi$ is dense in the space of all continuous functions defined on the half-space defined by the straight line. Thus, for example,

$$\overline{\text{span}}\{g\big((x-a)^2+y^2\big) \,:\, g \in C(\mathbb{R}),\, a \in \mathbb{R}\} = C(D)$$

where $D = \{(x,y) \,:\, y \geq 0\}$. In other words, one gets all even functions about the real axis.

(iii) If $\mathcal{A}$ is a set of straight lines in $\mathbb{R}^2$, then

$$\overline{\text{span}}\{g\big((x-a)^2+(y-b)^2\big) \,:\, g \in C(\mathbb{R}),\, (a,b) \in \mathcal{A}\} \neq C(\mathbb{R}^2)$$

if and only if all these lines have a common intersection point, and the angles between each of the lines is a rational multiple of π.

(iv) If $\mathcal{A}$ is an ellipse or parabola, then

$$\overline{\text{span}}\{g\big((x-a)^2+(y-b)^2\big) \,:\, g \in C(\mathbb{R}),\, (a,b) \in \mathcal{A}\} = C(\mathbb{R}^2)\,.$$

It would be interesting to determine necessary and sufficient conditions on the set $\mathcal{A}$ so that density holds.

§3. "Ridge" functions

Let us look at another simple set Φ. Let $h_1(\mathbf{x}), \ldots, h_m(\mathbf{x})$ be any m fixed continuous functions from $\mathbb{R}^n$ to $\mathbb{R}$. Let $\Phi = \text{span}\{h_1, \ldots, h_m\}$. In other words, consider

$$\mathcal{M} = \text{span}\{g\big(\sum_{i=1}^m a_i h_i(\mathbf{x})\big) : \mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{R}^m,\, g \in C(\mathbb{R})\}\,.$$

When is $\mathcal{M}$ dense in $C(\mathbb{R}^n)$?

a) The answer is quite simple, and since the proof is short and elegant, we give it here.

Proposition B. *$\overline{\mathcal{M}} = C(\mathbb{R}^n)$ if and only if $H(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_m(\mathbf{x}))$ separates points. That is, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{y}$, there exists $i \in \{1, \ldots, m\}$ such that $h_i(\mathbf{x}) \neq h_i(\mathbf{y})$.*

Proof: Consider the linear span of the set

$$\big(\sum_{i=1}^m a_i h_i(\mathbf{x})\big)^k$$

as we vary over all $\mathbf{a} \in \mathbb{R}^m$ and $k = 0, 1, 2, \ldots$. This is an algebra generated by

$$h_1^{\ell_1}(\mathbf{x}) \cdots h_m^{\ell_m}(\mathbf{x})$$

where the ℓ_i are non-negative integers. Furthermore this algebra contains the constant function and separates points. Thus the density follows from the Stone–Weierstrass Theorem. ■

b) Similar to a result of the previous section, assuming we have density in the above problem, it is not necessary to run over all vectors $\mathbf{a} \in \mathbb{R}^m$. It always suffices to run over all $\mathbf{a} \in \mathcal{A}$, where $\mathcal{A}$ is any subset of $\mathbb{R}^m$ for which no *non-trivial homogeneous polynomial* vanishes on $\mathcal{A}$. In general this sufficient condition on $\mathcal{A}$ is not necessary.

A *Ridge Function*, in its simplest sense, is a function of the form

$$f(\mathbf{x}) = g(\mathbf{a} \cdot \mathbf{x})$$

where $\mathbf{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ is fixed (a *direction*), and $\mathbf{a} \cdot \mathbf{x} = \sum_{i=1}^n a_i x_i$. That is, $m = n$ and $h_i(\mathbf{x}) = x_i$ in the previous example. (We will also consider a simple generalization thereof, namely

$$f(\mathbf{x}) = g(A\mathbf{x})$$

where A is a $k \times n$ matrix and $g : \mathbb{R}^k \to \mathbb{R}$.)

Ridge functions are constant on the hyperplanes $\mathbf{a} \cdot \mathbf{x} = c$ and thus are particularly simple functions. Ridge functions are used, with varying degrees of success, in different fields. In Partial Differential Equations they are called *Plane Waves*. (In general, linear combinations of ridge functions occur in the study of hyperbolic p.d.e.'s with constant coefficients.) In Statistics they are used in the theory of *Projection Pursuit* and *Projection Regression*. They are used in the theory of Computerized Tomography (the name Ridge Function was coined by Logan, Shepp [4] in one of the seminal papers on tomography). In Neural Networks they are used. (More specifically in a model in Neural Networks concerned with *Multilayer Feedforward Neural Networks* with *Input*, *Hidden* and *Output* layers.) Finally, approximation theorists are interested in using Ridge Functions as a method of approximating complicated (multivariate) functions by simple functions (linear combinations of Ridge Functions).

There is, unfortunately, not very much known about approximating using Ridge Functions. We mention here some of the results connected with the density problem.

c) What are necessary and sufficient conditions on a set $\mathcal{A} \subseteq \mathbb{R}^n$ such that

$$\mathcal{M}_{\mathcal{A}} = \operatorname{span}\{g(\mathbf{a} \cdot \mathbf{x}) : \mathbf{a} \in \mathcal{A},\ g \in C(\mathbb{R})\}$$

is dense in $C(\mathbb{R}^n)$?

For $n = 2$ it is known that it is both necessary and sufficient that $\mathcal{A}$ contains an infinite number of pairwise linearly independent points. (Note that if $\mathbf{a} \in \mathcal{A}$, then we add nothing to $\mathcal{M}_{\mathcal{A}}$ if we adjoin $\alpha\mathbf{a}$ to $\mathcal{A}$.) This result has been proved in the literature at least 3–4 times. The earliest reference we have found is to an article by Vostrecov, Kreines [6] (early 60's), where they in fact proved the following result.

Theorem C. $\overline{\mathcal{M}_{\mathcal{A}}} = C(\mathbb{R}^n)$ *if and only if no non-trivial homogeneous polynomial vanishes on* $\mathcal{A}$.

This result seems to have not been noticed, as very partial results were later reproved by others.

d) Let $\mathcal{A}$ be a subset of all $k \times n$ real matrices, $1 \le k < n$, fixed. Set

$$\mathcal{M}_{\mathcal{A}} = \operatorname{span}\{g(A\mathbf{x}) : A \in \mathcal{A},\, g \in C(\mathbb{R}^k)\}\,.$$

Necessary and sufficient conditions for when $\overline{\mathcal{M}_{\mathcal{A}}} = C(\mathbb{R}^n)$ are as follows (see Lin, Pinkus [3]):

For each $A \in \mathcal{A}$, let $L(A)$ denote the linear subspace spanned by the rows of A. (If $L(A) = L(B)$, then

$$\overline{\operatorname{span}}\{g(A\mathbf{x}) : g \in C(\mathbb{R}^k)\} = \overline{\operatorname{span}}\{g(B\mathbf{x}) : g \in C(\mathbb{R}^k)\}\,.)$$

Set

$$L(\mathcal{A}) = \bigcup_{A \in \mathcal{A}} L(A)\,.$$

(If $k = 1$, $L(A)$ is a line through the origin.)

Theorem D. $\mathcal{M}_{\mathcal{A}}$ *is dense in* $C(\mathbb{R}^n)$ *if and only if* $L(\mathcal{A})$ *is not contained in the zero set of any non-trivial polynomial. (Or homogeneous polynomial, since* $L(\mathcal{A})$ *is homogeneous.)*

Here are some facts connected with this result.

1. If $L(\mathcal{A})$ is not contained in the zero set of some non-trivial polynomial, then not only is $\mathcal{M}_{\mathcal{A}}$ dense in $C(\mathbb{R}^n)$, but in fact more is true. Namely, the polynomials are explicitly contained in the set $\mathcal{M}_{\mathcal{A}}$. (As finite linear combinations of $g(A\mathbf{x})$, g polynomial, $A \in \mathcal{A}$.)

2. If there exists a non-trivial polynomial p which vanishes on $L(\mathcal{A})$, then we can prove that for any $\phi \in C_0^\infty$, the function $\psi = p(D)\phi$ satisfies

$$\int g(A\mathbf{x})\psi(\mathbf{x})\,d\mathbf{x} = 0$$

for all $g \in C(\mathbb{R}^k)$ and for all $A \in \mathcal{A}$. (Here $p(D) = p(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n})$.) Thus we explicitly construct simple linear functionals which annihilate $\mathcal{M}_{\mathcal{A}}$. The density result therefore holds for any space where the Weierstrass Theorem holds, and the above is also a linear functional, for some $\phi \in C_0^\infty$.

3. If $\mathcal{A} = \mathcal{A}_1 \bigcup \mathcal{A}_2$, then $\mathcal{M}_{\mathcal{A}}$ is dense in $C(\mathbb{R}^n)$ if and only if $\mathcal{M}_{\mathcal{A}_i}$ is dense in $C(\mathbb{R}^n)$ for $i = 1$ and/or $i = 2$.

4. Of course, if $\mathcal{A}$ contains only a finite number of points (elements), then $\mathcal{M}_{\mathcal{A}}$ is not dense in $C(\mathbb{R}^n)$.

5. If $k = n - 1$, then $\mathcal{M}_{\mathcal{A}}$ is dense in $C(\mathbb{R}^n)$ if $\mathcal{A}$ contains an infinite number of pairwise "distinct" matrices A of full rank $n - 1$. (Distinct here means that $L(A) \neq L(B)$.)

Finally we note some related result. (See Lin, Pinkus [3].)

e) What is $\overline{\mathcal{M}_{\mathcal{A}}}$ in general?

$$\begin{aligned}\overline{\mathcal{M}_{\mathcal{A}}} &= \overline{\text{span}}\{g(A\mathbf{x}) : A \in \mathcal{A}, g \in C(\mathbb{R}^k)\} \\ &= \overline{\text{span}}\{q(\mathbf{x}) : q \text{ polynomial, } p(D)q = 0 \text{ for every} \\ &\qquad \text{polynomial which vanishes on } L(\mathcal{A})\}.\end{aligned}$$

This is related to the theory of *Polynomial Ideals*. (The set of polynomials which vanish on $L(\mathcal{A})$ is a Polynomial Ideal (of a fairly simple form).) As such we could actually replace "$p(D)q = 0$ for every ....." by a finite number of such p.

f) "Variable" Directions.

Assume we are given a fixed positive integer r. For $\mathcal{A} = \{A_1, \ldots, A_r\}$, where the A_i are some $k \times n$ matrices, we have that $\mathcal{M}_{\mathcal{A}}$ is not dense in $C(\mathbb{R}^n)$. What if we keep r fixed, but vary the "directions" $A_1, \ldots, A_r$ (as well as the g_i)? That is, what if we approximate from the highly nonlinear set

$$\{\sum_{i=1}^{r} g_i(A_i\mathbf{x}) : g_i \in C(\mathbb{R}^k), A_i \ k \times n \text{ matrix}\}\,?$$

Unfortunately, one can prove that density does not hold for any r (on any compact set with interior).

Acknowledgments. This research was supported by the fund for the promotion of research at the Technion.

References

1. Chui, C. K. and Xin Li, Approximation by ridge functions and neural networks with one hidden layer, *J. Approx. Theory* **70** (1992), 131–141.
2. Leshno, L., V. Ya. Lin, A. Pinkus, and S. Schocken, Multilayer feedforward networks with a non-polynomial activation function can approximate any function, *Neural Networks*, to appear.
3. Lin, V. Ya. and A. Pinkus, Fundamentality of Ridge Functions, *J. Approx. Theory*, to appear.
4. Logan, B. F. and L. A. Shepp, Optimal reconstruction of a function from its projections, *Duke Math. J.* **42** (1975), 645–659.
5. Vitushkin, A. G. and G. M. Henkin, Linear superpositions of functions, *Uspehi Mat. Nauk* **22** (1967), 77–124 = *Russian Math. Surveys* **22** (1967), 77–125.
6. Vostrecov, B. A. and M. A. Kreines, Approximation of continuous functions by superpositions of plane waves, *Dokl. Akad. Nauk SSSR* **140** (1961), 1237–1240 = *Soviet Math. Dokl.* **2** (1961), 1326–1329.

Vladimir Ya. Lin
Department of Mathematics
Technion, I. I. T.
Haifa, 32000
Israel

Allan Pinkus
Department of Mathematics
Technion, I. I. T.
Haifa, 32000
Israel