The General Approximation Theorem

Donald C. Wunsch
Missouri University of Science and Technology, dwunsch@mst.edu

Alexander N. Gorban

THE GENERAL APPROXIMATION THEOREM

A. N. Gorban'
Krasnoyarsk Computing Center, Russian Academy of Sciences, Siberian Branch
Krasnoyarsk-36, 660036, Russian Federation
E-mail: gorban@cc.krascience.rssi.ru

Donald C. Wunsch II
Applied Computational Intelligence Laboratory
Department of Electrical Engineering, Texas Tech University,
Lubbock, TX 79409-3102, E-mail: Dwunsch@aol.com
http://www.acil.ttu.edu

Abstract

A general approximation theorem is proved. It uniformly envelopes both the classical Stone theorem and approximation of functions of several variables by means of superpositions and linear combinations of functions of one variable. This theorem is interpreted as a statement on universal approximating possibilities ("approximating omnipotence") of arbitrary nonlinearity. For the neural networks, our result states that the function of neuron activation must be nonlinear – and nothing else.

Keywords — Approximation, Superposition, Neural networks, Stone-Weierstrass theorem.

1. INTRODUCTION

The question of representing continuous functions of several variables by superposition of continuous functions with fewer variables has been the essence of Hilbert's 13th problem.

In [1], Kolmogorov proved an elegant theorem: *every continuous function of n variables defined in the standard n-dimensional cube can be represented in the following form:*

$$f(x_1, x_2, \ldots, x_n) = \sum_{q=1}^{2n+1} h_q \left[\sum_{p=1}^{n} \varphi_q^p (x_p) \right],$$

where the functions $h_q(u)$ *and* $\varphi_q^p(x_p)$ *are continuous, and moreover, the functions* $\varphi_q^p(x_p)$ *are standard, i.e. they are independent of the function f.*

It should be noted that the functions $\varphi_q^p(x_p)$ used here are essentially non-smooth and very exotic.

The Kolmogorov theorem is often cited in papers and books on neural networks, though, as it is easy to see, it bears no relation to the latter for the following reasons:

1) by means of neural networks *approximations* of functions of several variables are constructed, while the Kolmogorov theorem substantiates the possibility of their *exact* representation;

2) nonlinear elements used in neural networks can be practically arbitrary, and usually they calculate either smooth or piecewise linear functions, while in the Kolmogorov theorem very specific nonsmooth functions are employed.

Thus, the Kolmogorov theorem is devoted to representation of functions of several variables through very specific functions of one variable, but neural networks allow approximation of functions of several variables through practically arbitrary nonlinear function of one variable.

The possibility of uniform approximation of continuous functions through polynomials is proved in the classic Weierstrass theorem. A strong generalization of the Weierstrass theorem is the Stone theorem [2]:

Let consider a compact space X and algebra C(X) of continuous functions on X with real values. If $E \subseteq C(X)$ is a closed subalgebra in C(X), $1 \in E$ and the functions from E separate the points in E (i.e. for any two different $x, y \in X$ exists such a function $g \in E$ that $g(x) \neq g(y)$), then E=C(X).

The Stone theorem generalizes the Weierstrass theorem in two directions. First, functions on arbitrary compact are considered rather than functions of several variables only. Second, the following statement was proved, which was new even for the functions of one variable: not only the set of polynomials of coordinate functions is complete, but, in general, a ring of polynomials of any set of functions separating the points.

Therefore, the set of trigonometric polynomials are complete, and the set of linear combinations of functions of the form $\exp(-(x-x_0, Q(x-x_0)))$, where $(x, Q(x))$ is a positive definite quadratic form, e.g.

In this paper, we study, approximate representations of functions of several variables by functions of one variable. In a contrast to the Kolmogorov theorem, the question we address is: how broad is the class of functions that can be approximated using a single, ***arbitrarily taken***, and not specially constructed, nonlinear continuous function?

The answer is: *every continuous function can be arbitrarily accurately approximated by operations of addition, multiplication by a number, and superposition of an arbitrary number (one is sufficient) of continuous nonlinear functions of one variable.*

Renewed interest in the classical question of approximation of functions of several variables by superpositions and sums of functions of one variable and a new version of this question (confined to one arbitrarily taken nonlinear function) have been invoked by neurocomputing studies.

The question what functions they are able to approximate is becoming topical. Relevant theorems on completeness for several versions of the neural networks have been proved [3-6]. They are distinct in admissible architectures of the networks, in functions that are computed by an individual "neuron", etc. The present work proves the theorem on completeness ***for arbitrary continuous functions***.

For the neural networks, our result states that the function of neuron activation must be nonlinear – and nothing else. Whatever this nonlinearity is, the network of connections can be constructed, and coefficients of linear connections between the neurons can be adjusted in such a way that the neural network will compute any continuous function from its input signals with any given accuracy.

2. SEMIGROUPS OF CONTINUOUS FUNCTIONS OF ONE VARIABLE

Consideration of functions of one variable is necessary for further study of functions of many variables. On the other hand, the theorem on density of any semigroup of continuous functions which includes at least one nonlinear function is also of independent interest.

Let us consider the space $C(R)$ of continuous functions on a real axis in the topology of the uniform convergence on compact sets. The space $C(R)$ with superposition of functions $(f \circ g)(x) = f(g(x))$ on it is a semigroup. Function $\mathrm{id}(\mathrm{id}(x) \equiv x)$ is a unit in this semigroup.

Theorem 1. *Let E be a closed subspace in C(R) which is a semigroup, 1∈E and* id∈*E (1 is a function identically equal to 1). Then, either E=C(R) or E is a subspace of linear functions (f(x)=ax+b).*

The proof is based on several lemmas.

Lemma 1. *Under the conditions of Theorem 1, let there exist a function f∈E which is not linear. Then, there exists a twice continuously differentiable function* g∈*E which is not linear.*

Lemma 2. *Under the conditions of Theorem 1, let there exist a twice differentiable function g∈E which is not linear. Then, the function $q(x)=x^2$ is in E.*

Lemma 3. *Let under the conditions of the Theorem 1 the function $q(x)=x^2$ be in E. Then, E is a ring: for every f,g∈E their product fg∈E.*

From the lemmas 1-3 it follows that under the conditions of the Theorem 1, if E has even one nonlinear function then E is a ring, and contains, in particular, all polynomials. Hence, by the Weierstrass theorem, it follows that $E= C(R)$.

3. GENERAL APPROXIMATION THEOREM FOR FUNCTIONS OF SEVERAL VARIABLES

Consider a compact space X and algebra $C(X)$ of continuous real functions on X. Let $E \in C(X)$ be a linear space, $C(R)$ be a space of continuous functions on the real axis R, $f \in C(R)$ be a nonlinear function and for any $g \in E$, $f(g) \in E$ holds. In this case let us say that E is closed with respect to nonlinear unary operation f.

Example: *a set of functions of n variables, which can be exactly represented using the given function f of one* *variable, linear functions and superposition operation is a linear space, closed with respect to nonlinear unary operation f. The closure of this set in the space of all continuous functions possesses the same property.*

Remark. *The linear space* $E \in C(X)$ *is closed with respect to nonlinear operation* $f(x) = x^2$ *if and only if E is a ring.* Indeed, $fg = \frac{1}{2}\left[(f+g)^2 - f^2 - g^2\right]$, therefore for the linear space $E \in C(X)$ the closedness with respect to the unary operation $f(x) = x^2$ is equivalent to closedness with respect to the product of functions.

According to the above remark, the Stone theorem can be reformulated as follows. *Let* $E \in C(X)$ *be closed linear subspace in* $C(X)$, $1 \in E$, *the functions from E separate points of X and E be closed with respect to nonlunear unary operation* $f(x) = x^2$. *Then* $E = C(X)$.

Our generalization of the Stone theorem consists in change of $f(x) = x^2$ with arbitrary nonlinear continuous function.

Theorem 1. *Let* $E \in C(X)$ *be closed linear subspace in* $C(X)$, $1 \in E$, *the functions from E separate points of X and E be closed with respect to nonlinear unary operation* $f \in C(R)$. *Then* $E = C(X)$.

Proof. Consider the set of all such $p \in C(R)$ that $p(E) \subseteq E$, i.e. for any $g \in E$ holds $p(g) \in E$. Let us denote this set P_E. It possesses the following properties:

1) P_E is a semigroup with respect to superposition of functions;

2) P_E is closed linear subspace in $C(R)$ (in the uniform convergence topology on compacts);

3) $1 \in P_E$ and $\mathrm{id} \in P_E$ $(\mathrm{id}(x) = x)$;

4) P_E includes at least one continuous nonlinear function.

The rest of the proof follows from Theorem 2, which in our work is, essentially, a preparation theorem on semigroups of functions.

4. ALGEBRAIC VARIANT OF APPROXIMATION THEOREM

The return to the classic problem on representation of functions of several variables through superpositions and sums of functions of one variable is connected with investigations of neural networks. There are two, not one, classic problems, and only the second of them directly relates to neural networks:

1. Is it possible to obtain **exact representation** of a function of several variables by means of superposition of functions of smaller number of variables?

2. Is it possible to obtain **arbitrarily close approximation** of a function of several variables by means of some simpler functions and operations?

In the present section, the central place belongs to a theorem which is similar in form to the generalized approximation theorem, but relates to the first rather than the second problem, since it states the possibility of exact representation of all polynomials of several variables by means of arbitrary nonlinear polynomial of one variable, linear functions and superposition operations. The distance between the first and the second problems appears to be not so large.

Let $R[X]$ be a ring of polynomials of one variable over a field R of characteristics 0, and $E \subset R[X]$ be a linear space of polynomials over R. The following simple proposition is an algebraic analogue of Theorem 1.

Proposition 1. *If E is closed with respect to superposition of polynomials, and contains all the polynomials of the first power and at least one polynomial* $p(x)$ *of the power* $m > 1$, *then* $E = R[X]$.

Denote $R[X_1, \ldots, X_n]$ the ring of polynomials of n variables over field R of characteristic 0.

For any linear subspace $E \subseteq R[X_1, \ldots, X_n]$ consider the set of algebraic unary operations which transform elements of E into elements of E:

$$P_E = \{p \in R[X] \mid p(g(x_1, \ldots, x_n)) \in E \text{ for any } g(x_1, \ldots, x_n) \in E\}.$$

Proposition 2. *For any linear subspace* $E \subseteq R[X_1, \ldots, X_n]$, *the set of polynomials* P_E *is a linear space over R, closed with respect to superposition and contains all uniform polynomials of first power.*

Proposition 3. *If a linear subspace* $E \subseteq R[X_1, \ldots, X_n]$ *contains* 1 *and* P_E *includes at least one polynomial of degree* $m > 1$ *(i.e. nonlinear), then* $P_E = R[X]$.

Theorem 3. *Let* $E \subseteq R[X_1, \ldots, X_n]$ *be linear subspace in* $R[X_1, \ldots, X_n]$, *E contains all the polynomials of the first power and is closed with respect to nonlinear unary operation* $p \in R[X]$. *Then* $E = R[X_1, \ldots, X_n]$.

Proof follows from Propositions 2, 3 and the fact that a ring of polynomials that includes all the polynomials of first power coincides with $R[X_1, \ldots, X_n]$.

Let p be a polynomial of one variable, $E_p[X_1, \ldots, X_n]$ be set of polynomials of n variables, which can be obtained from p and polynomials of first power belonging to $R[X_1, \ldots, X_n]$ by means of operations of superposition, addition and multiplication by number.

The following two propositions give a convenient characterization of $E_p[X_1, \ldots, X_n]$ and follow directly from definitions.

Proposition 4. *The set* $E_p[X_1, \ldots, X_n]$ *is a linear space over R and for any polynomial* $g(x_1, \ldots, x_n)$ *from* $E_p[X_1, \ldots, X_n]$

$$p(g(x_1, \ldots, x_n)) \in E_p[X_1, \ldots, X_n].$$

Proposition 5. *For given p the family of linear subspaces* $L \subseteq R[X_1, \ldots, X_n]$ *containing all the polynomials of first power and satisfying the condition*

$$\text{if } g(x_1, \ldots, x_n) \in L, \text{ then } p(g(x_1, \ldots, x_n)) \in L$$

is closed with respect to intersections. The minimal inclusion element of this family is $E_p[X_1, \ldots, X_n]$.

From Propositions 4, 5 and Theorem 3 we obtain the following statement.

Corollary. *For any polynomial p of power* $m > 1$

$$E_p[X_1, \ldots, X_n] = R[X_1, \ldots, X_n].$$

Thus, from p and polynomials of the first power by means of operations of superposition, addition and multiplication by a number it is possible to obtain all the elements of $R[X_1, \ldots, X_n]$.

5. DISCUSSION

Investigation of neural networks has complemented Weierstrass and Stone theorems. In addition, the theorem on approximation of functions of several variables is valid: every continuous function of several variables can be approximated arbitrarily accurately using linear functions, superposition operation and an arbitrary function of one variable.

When we can use superposition of functions, linear functions and at least one arbitrary non-linear continuous function of one variable we can approximate every continuous function of several variables.

These theorems can be interpreted as statements about universal approximation properties of every nonlinearity: linear operations and cascade combinations can be used to produce from arbitrary nonlinear elements every required results with preassigned accuracy.

The theorem proved uniformly envelopes both the classical Stone theorem and approximation of functions of several variables by means of superpositions and linear combinations of functions of one variable.

REFERENCES

1. Kolmogorov A.N. On representation of continuous functions of several variables as superpositions of continuous functions of one variable. Dokl. AN SSSR, 1957. V. 114. No 5. P. 953-956.
2. Stone M.N. The generalized Weierstrass approximation theorem. Math. Mag., 1948. V. 21. P. 167-183, 237-254.
3. Cybenko G. Approximation by superposition of a sigmoidal function. Mathematics of Control, Signals, and Systems, 1989. Vol. 2. P. 303-314.
4. Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators. Neural Networks. 1989. Vol. 2. P. 359-366.
5. Kochenov D.A., Rossiev D.A. Approximations of functions of $C(A, B)$ class by neural-net predictors (architectures and results). AMSE Transaction, Scientific Siberian, A. 1993, Vol. 6. Neurocomputing. P. 189-203. Tassin, France.
6. Gilev S.E., Gorban A.N. On completeness of the class of functions computable by neural networks. Proc. of the World Congress on Neural Networks (WCNN'96). Sept. 15-18, 1996, San Diego, CA, Lawrens Erlbaum Accociates, 1996. P. 984-991.