NOTE Communicated by Halbert White

Representation Properties of Networks: Kolmogorov's Theorem Is Irrelevant

Federico Girosi
Tomaso Poggio
Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA 02142 USA
and
Center for Biological Information Processing, Whitaker College, Cambridge, MA 02142 USA

Many neural networks can be regarded as attempting to approximate a multivariate function in terms of one-input one-output units. This note considers the problem of an exact representation of nonlinear mappings in terms of simpler functions of fewer variables. We review Kolmogorov's theorem on the representation of functions of several variables in terms of functions of one variable and show that it is irrelevant in the context of networks for learning.

1 Kolmogorov's Theorem: An Exact Representation Is Hopeless

A crucial point in approximation theory is the choice of the representation of the approximant function. Since each representation can be mapped in an appropriate network choosing the representation is equivalent to choosing a particular network architecture. In recent years it has been suggested that a result of Kolmogorov (1957) could be used to justify the use of multilayer networks composed of simple one-input–one-output units. This theorem and a previous result of Arnol'd (1957) can be considered as the definitive disproof of *Hilbert's conjecture* (his thirteenth problem, Hilbert 1900): *there are continuous functions of three variables, not representable as superpositions of continuous functions of two variables.*

The original statement of Kolmogorov's theorem is the following (Lorentz 1976):

Theorem 1.1. (Kolmogorov 1957). *There exist fixed increasing continuous functions $h_{pq}(x)$, on $I = [0,1]$ so that each continuous function f on I^n can be written in the form*

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} g_q\left(\sum_{p=1}^{n} h_{pq}(x_p)\right),$$

where g_q are properly chosen continuous functions of one variable.

Neural Computation **1, 465–469** (1989)

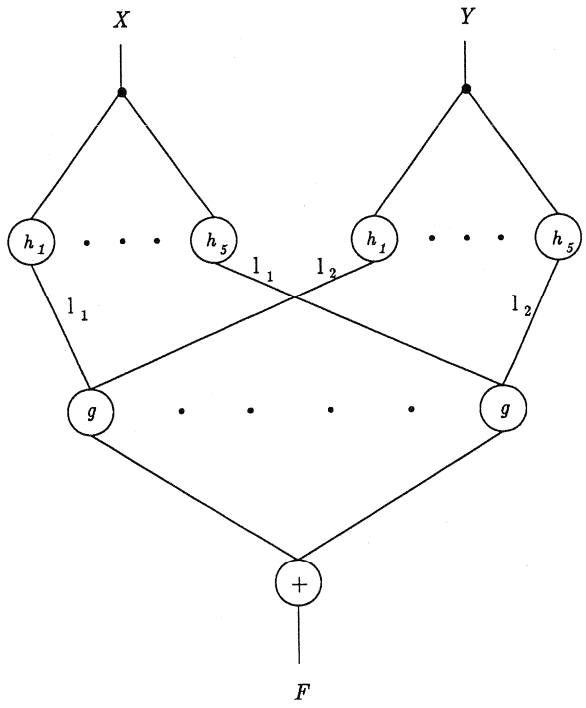

Figure 1: The network representation of an improved version of Kolmogorov's theorem, due to Kahane (1975). The figure shows the case of a bivariate function. The Kahane's representation formula is $f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} g[\sum_{p=1}^{n} l_p h_q(x_p)]$ where h_q are strictly monotonic functions and l_p are strictly positive constants smaller than 1.

This result asserts that every multivariate continuous function can be represented by the superposition of a small number of univariate continuous functions. In terms of networks this means that every continuous function of many variables can be computed by a network with two hidden layers (see Figure 1) whose hidden units compute continuous functions (the functions g_q and h_{pq}).

Does Kolmogorov's theorem, in its present form, prove that a network with two hidden layers is a good and usable representation? The answer is definitely no. There are at least two reasons for this:

1. In a network implementation that has to be used for learning and generalization, some degree of smoothness is required for the func-

tions corresponding to the units in the network. Smoothness of the h_{pq} and of the g_q is important because the representation must be smooth in order to generalize and be stable against noise. A number of results of Vituskin (1954, 1977) and Henkin (1964) show, however, that the inner functions h_{pq} of the Kolmogorov's theorem are highly not smooth (they can be regarded as "hashing" functions). Due to this "wild" behavior of the inner functions h_{pq}, the functions g_q do not need to be smooth, even for differentiable functions f (de Boor 1987).

2. Useful representations for approximation and learning are *parametrized* representations that correspond to networks with fixed units and modifiable parameters. Kolmogorov's network is not of this type: the form of g_q (corresponding to units in the second "hidden" layer) depends on the specific function f to be represented (the h_{pq} are independent of it). g_q is at least as complex, for instance in terms of bits needed to represent it, as f.

A stable and usable *exact* representation of a function in terms of two or more layers network seems hopeless. In fact the result obtained by Kolmogorov can be considered as a "pathology" of the continuous functions: it fails to be true if the inner functions h_{pq} are required to be smooth, as it has been shown by Vitushkin (1954). The theorem, though mathematically surprising and beautiful, cannot be used by itself in any constructive way in the context of networks for learning. This conclusion seems to echo what Lorentz (1962) wrote, more than 20 years ago, asking "Will it [Kolmogorov's theorem] have useful applications?... One wonders whether Kolmogorov's theorem can be used to obtain positive results of greater [than trivial] depth." Notice that this leaves open the possibility of finding good and well founded approximate representations. This argument is discussed in some length in Poggio and Girosi (1989), and a number of results have been recently obtained by some authors (Hornik *et al.* 1989; Stinchcombe and White 1989; Carroll and Dickinson 1989; Cybenko 1989; Funahashi 1989; Hecht-Nielsen 1989).

The next section reviews Vitushkin's main results.

2 The Theorems of Vitushkin

The interpretation of Kolmogorov's theorem in term of networks is very appealing: the representation of a function requires a fixed number of nodes, polynomially increasing with the dimension of the input space. Unfortunately, these results are somewhat pathological and their practical implications very limited. The problem lies in the inner functions of Kolmogorov's formula: although they are continuous, theorems of Vitushkin and Henkin (Vitushkin 1964, 1977; Henkin 1964; Vitushkin and Henkin 1967) prove that they must be highly nonsmooth. One could ask if it is

possible to find a superposition scheme in which the functions involved are smooth. The answer is negative, even for two variable functions, and was given by Vitushkin with the following theorem (1954):

Theorem 2.1. (Vitushkin 1954). *There are $r(r = 1, 2, \ldots)$ times continuously differentiable functions of $n \geq 2$ variables, not representable by superposition of r times continuously differentiable functions of less than n variables; there are r times continuously differentiable functions of two variables that are not representable by sums and continuously differentiable functions of one variable.*

We notice that the intuition underlying Hilbert's conjecture and theorem 2.1 is the same: not all the functions with a given degree of complexity can be represented in simple way by means of functions with a lower degree of complexity. The reason for the failing of Hilbert's conjecture is a "wrong" definition of complexity: Kolmogorov's theorem shows that the number of variables is not sufficient to characterize the complexity of a function. Vitushkin showed that such a characterization is possible and gave an explicit formula. Let f be an r times continuously differentiable function defined on I^n with all its partial derivatives of order r belonging to the class $Lip[0,1]^\alpha$. Vitushkin puts $\chi = (r + \alpha)/n$ and shows that it can be used to measure the inverse of the complexity of a class of functions. In fact he succeded in proving the following:

Theorem 2.2. (Vitushkin 1954). *Not all functions of a given characteristic $\chi_0 = q_0/k_0 > 0$ can be represented by superpositions of functions of characteristic $\chi = q/k > \chi_0, q \geq 1$.*

Theorem 2.1 is easily derived from this result.

Acknowledgments

We acknowledge support from the Defense Advanced Research Projects Agency under contract number N00014-89-J-3139. Tomaso Poggio is supported by the Uncas & Helen Whitaker Chair at MIT.

References

Arnol'd, V. I. 1957. On functions of three variables. *Dokl. Akad. Nauk SSSR* **114**, 679–681.

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

Cybenko, G. 1989. Approximation by superposition of a sigmoidal function. *Math. Control Systems Signals*, in press.

de Boor, C. 1987. Multivariate approximation. In *The State of the Art in Numerical Analysis*, A. Iserles and M. J. D. Powell, eds., pp. 87–109. Clarendon Press, Oxford.

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

Hecht-Nielsen, R. 1989. Theory of backpropagation neural network. In *Proceedings of the International Joint Conference on Neural Networks*, pp. I-593–I-605, Washington D.C., June 1989. IEEE TAB Neural Network Committee.

Henkin, G. M. 1964. Linear superpositions of continuously differentiable functions. *Dokl. Akad. Nauk SSSR* **157**, 288–290.

Hilbert, D. 1900. Mathematische probleme. *Nachr. Akad. Wiss. Göttingen*, 290–329.

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

Kahane, J. P. 1975. Sur le theoreme de superposition de Kolmogorov. *J. Approx. Theory* **13**, 229–234.

Kolmogorov, A. N. 1957. On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR* **114**, 953–956.

Lorentz, G. G. 1976. On the 13-th problem of Hilbert. In *Proceedings of Symposia in Pure Mathematics*, pp. 419–429, Providence, RI, 1976. American Mathematical Society.

Lorentz, G. G. 1962. Metric entropy, widths, and superposition of functions. *Am. Math. Monthly* **69**, 469–485.

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

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

Vitushkin, A. G. 1954. On Hilbert's thirteenth problem. *Dokl. Akad. Nauk SSSR* **95**, 701–704.

Vitushkin, A. G. 1964. Some properties of linear superposition of smooth functions. *Dokl. Akad. Nauk SSSR* **156**: 1003–1006.

Vitushkin, A. G. 1977. *On Representation of Functions by Means of Superpositions and Related Topics*. L'Enseignement Mathematique.

Vitushkin, A. G., and Henkin, G. M. 1967. Linear superposition of functions. *Russian Math. Surveys* **22**, 77–125.

Received 17 July 1989; accepted 30 August 1989.