Anlauf, J. K., and Biehl, M. 1989. The AdaTron: An adaptive perceptron algorithm. *Europhys. Lett.* **10**, 687.

Anlauf, J. K., and Biehl, M. 1990. Properties of an adaptive perceptron learning algorithm. In *Parallel Processing in Neural Systems and Computers*, Eckmiller, R., Hartmann, G. and Hauske, G., eds., Elsevier. This work presents a *parallel* algorithm, which converges to a local minimum of an energy function when no solution exists.

Cover, T. 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Elect. Comput.* **14**, 326.

Frean, M. 1990. The upstart algorithm: A method for constructing and training feedforward neural networks. *Neural Comp.* **2**, 198.

Gardner, E. 1988. The space of interactions in neural network models. *J. Phys. A* **21**, 257.

Grossman, T., Meir, R., and Domany, E. 1989. Learning by choice of internal representations. *Complex Syst.* **2**, 555.

Ho, E., and Kashyap, R. L. 1965. An algorithm for linear inequalities and its applications. *IEEE Trans. Elect. Comp.* **EC-14**, 683.

Karmarkar, N. 1984. A new polynomial time algorithm for linear programming. *Combinatorica* **4**, 373–395.

Khachian, L. G. 1979. A polynomial time algorithm for linear programming. *Dokl. Akad. Nauk USSR* **244**(5), 1093–1096. Translated in *Soviet Math. Dokl.* **20**, 191–194.

Kinzel, W., and Rujan, P. 1990. Improving a network generalization ability by selecting examples. *Europhys. Lett.* **13**, 473.

Krauth, W., and Mezard, M. 1987. Learning algorithms with optimal stability in neural networks. *J. Phys. A* **20**, L745.

Mezard, M., and Nadal, J. 1989. Learning in feedforward neural networks: The tiling algorithm. *J. Phys. A* **22**, 2191.

Muroga, S., Toda, I., and Takasu, S. 1961. Theory of majority decision elements. *J. Franklin Inst.* **271**, 376.

Nabutovsky, D., Grossman, T., and Domany, E. 1990. Learning by CHIR without storing internal representations. *Complex Syst.* **4**, 519.

Opper, M., Kinzel, W., Kleinz, J., and Niehl, R. 1990. On the ability of the optimal perceptron to generalise. Preprint.

Rosenblatt, F. 1962. *Principles of Neurodynamics*. Spartan Books, New York.

Received 3 December 1990; accepted 1 July 1991.

Communicated by Halbert White

Kolmogorov's Theorem Is Relevant

Věra Kůrková
Institute of Computer Science, Czechoslovak Academy of Sciences, P. O. Box 5, 182 07 Prague 8, Czechoslovakia

We show that Kolmogorov's theorem on representations of continuous functions of n-variables by sums and superpositions of continuous functions of one variable is relevant in the context of neural networks. We give a version of this theorem with all of the one-variable functions approximated arbitrarily well by linear combinations of compositions of affine functions with some given sigmoidal function. We derive an upper estimate of the number of hidden units.

Hecht-Nielsen (1987) suggested that a remarkable mathematical result of Kolmogorov (1957) could provide new insights and tools for understanding multilayer neural networks. There are several theorems in different branches of mathematics named after this great Russian mathematician. The one mentioned by Hecht-Nielsen was a theorem disproving Hilbert's conjecture formulated as the thirteenth of the famous list of 23 open problems that Hilbert supposed to be of the greatest importance for the development of mathematics in this century.

The thirteenth problem, although formulated as a concrete minor hypothesis, is connected with the basic problem of algebra — the solution of polynomial equations. Could roots of a general algebraic equation of higher degree be expressed, analogously to the solution by radicals, by sums and compositions of a one-variable function of some suitable type? Hilbert conjectured that some continuous functions of three variables are not representable by sums and superpositions even of functions of two variables. This was refuted by Arnold (1956). Kolmogorov (1957) even proved a general representation theorem stating that any continuous function f defined on an n-dimensional cube is representable by sums and superpositions of continuous functions of only one variable. Kolmogorov's formula

$$f(x_1,\dots,x_n) = \sum_{q=1}^{2n+1} \varphi_q \left[\sum_{p=1}^{n} \psi_{pq}(x_p)\right] \tag{1.1}$$

readily brings to mind perceptron type networks with the qualification that the one-variable functions $\varphi_q (q = 1,\dots,2n+1)$ and ψ_{pq} $(p = 1,\dots,n,$ $q = 1,\dots,2n+1)$ are far from being any of the type of functions currently used in neurocomputing. In fact, having even fractal graphs, they are highly nonsmooth.

This was the reason for Girosi and Poggio's (1989) criticism of Hecht-Nielsen's proposal. They formulated two main reservations:

1. The functions ψ_{pq} are highly nonsmooth.
2. The functions φ_q depend on the specific function f and hence are not representable in a parameterized form.

We shall show that by replacing the equality in equation 1.1 by only an approximation, we can eliminate both of these difficulties. Highly nonsmooth functions encountered in mathematics are mostly constructed as limits or sums of infinite series of smooth functions. This is the case, e.g., with the classical Weierstrass's function with no derivative at any point and many other famous examples of functions with fractal graphs. Since in the context of neural networks we are interested only in approximations of functions, the only problem concerning the possible relevance of Kolmogorov's theorem for neurocomputing is whether Kolmogorov's construction can be modified in such a way that all of the one-variable functions are limits of sequences of smooth functions used in perceptron type networks.

By a perceptron type network we mean a multilayer network where units in each hidden layer sum up weighted inputs from the preceding layer, add to this sum a constant (bias), and then apply a sigmoidal nonlinearity, while units in the output layer sum only weighted inputs. So functions used in perceptron type networks are finite linear combinations of compositions of affine transformations of the real line E_1 with some given sigmoidal function [a function $\sigma : E_1 \rightarrow [0,1]$ with $\lim_{t\rightarrow-\infty} \sigma(t) = 0$ and $\lim_{t\rightarrow\infty} \sigma(t) = 1$]. We call them staircase-like functions of a sigmoidal type (or of a type σ).

Kolmogorov's construction of the functions φ_q and ψ_{pq} and their later improvements by Lorentz (1962) and Sprecher (1965) are, in fact, perfectly suited for staircase-like functions of any sigmoidal type. Being very complex, all of these arguments contain a lot of unnecessary assumptions. But the only really relevant property of the functions used in inductive construction of one-variable functions φ_q and ψ_{pq} is that they have prescribed values on finitely many closed intervals; elsewhere they can be arbitrary, provided they are sufficiently bounded. However, such functions can be approximated arbitrarily well by staircase-like functions of any sigmoidal type (Kůrková 1991).

To illustrate the idea of Kolmogorov's construction of functions ψ_{pq}, recall the classical Devil's staircase (Fig. 1). Kolmogorov, probably inspired by this nineteenth-century construction, developed "the second generation Devil's staircase," something Mandelbrot (1982) would appreciate, by replacing in each induction step the already constructed Devil's staircase's steps (within a very small neighborhood of each) by smaller steps.

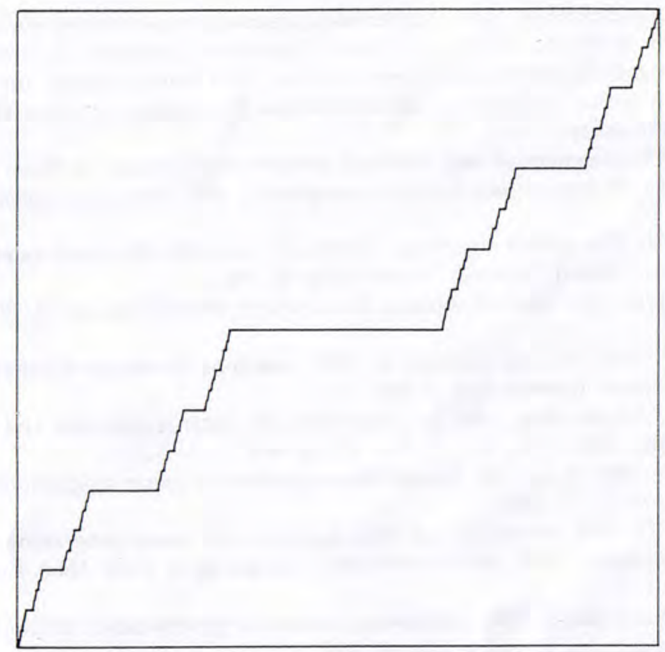

Figure 1: Devil's staircase.

The result was a strictly increasing function with, in contrast to the rectifiable classical Devil's staircase, a fractal graph. Nevertheless, both first and second generation Devil's staircases are limits of uniformly converging series of staircase-like functions of any sigmoidal type.

In contrast to the functions ψ_{pq}, being for the given dimension n universal, the functions φ_q depend on f. However, they can be also constructed as limits of staircase-like functions of any sigmoidal type. Consider for staircase-like functions ψ_p of any sigmoidal type, the function Ψ defined on the n-dimensional cube by $\Psi(x_1, \ldots, x_n) = \sum_{p=1}^{n} \psi_p(x_p)$. Ψ defines on the cube a Rubik's cube-like structure with small boxes having edges corresponding to the steps of ψ_p and gaps corresponding to the slopes of ψ_p. Suppose that the small boxes are mapped by Ψ into closed mutually disjoint subintervals of the real line. Ascribing to these intervals values of f at chosen points in the small boxes that Ψ maps into these intervals, we define a finite family of steps that can be approximated arbitrarily well by a staircase-like function φ of a given sigmoidal type. This function φ is representable in a parameterized form with the values of parameters depending on f. The function $\varphi \cdot \Psi$ approximates f on the subset of the cube formed by the union of all small boxes. The smaller the steps of ψ_p, the better the approximation. However, f is not approximated on the gaps. Now, we come to the reason why there are $2n+1$ terms under the summation in (1). By suitable shifts of the slopes of the staircase we can gain $2n+1$ Rubik's cube-like structures on the unit cube covering the n-dimensional cube sufficiently well in such a way that for each point there are more structures containing it in a box than structures containing it in a gap. We need $2n+1$ such structures, since at some point of the cube it may happen that each of its n coordinates is contained in the gaps of a different structure (at most n).

These are, roughly speaking, the ideas behind the proofs of the following theorems.

Theorem 1. *(Kůrková 1991). Let n, m be natural numbers with $n \geq 2$, $m \geq 2n+1$, and $\sigma : E_1 \to [0, 1]$ be any sigmoidal function. Then there exist such real numbers $w_{pq}(p = 1, \ldots, n, q = 1, \ldots, m)$ and functions $\psi_q(q = 1, \ldots, m)$ being limits of uniformly converging sequences of staircase-like functions of a type σ that for every continuous function $f : [0, 1]^n \to E_1$ there exists a continuous function $\varphi : E_1 \to E_1$ being a limit of a uniformly converging sequence of staircase-like functions of a type σ, such that for every $(x_1, \ldots, x_n) \in [0, 1]^n$*

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{m} \varphi \left[\sum_{p=1}^{n} w_{pq} \psi_q(x_p) \right]$$

Theorem 2. *(Kůrková 1991). Let $n \geq 2$ be a natural number, $\sigma : E_1 \to [0, 1]$ be a sigmoidal function, $f : [0, 1]^n \to E_1$ be a continuous function and ϵ a positive real number. Then there exist a natural number k and staircase-like functions of a type σ $\psi_{pi}, \varphi_i(i = 1, \ldots, k, p = 1, \ldots, n)$ such that for every $(x_1, \ldots, x_n) \in [0, 1]^n$*

$$\left| f(x_1, \ldots, x_n) - \sum_{i=1}^{k} \varphi_i \left[\sum_{p=1}^{n} \psi_{pi}(x_p) \right] \right| < \epsilon$$

Theorem 2 implies that any continuous function can be approximated arbitrarily well by a four-layer perceptron type network. However, several recent results (Funahashi 1989; Hecht-Nielsen 1989; Hornik *et al.* 1989; Cybenko 1989; Carroll and Dickinson 1989; Stinchcombe and White 1989, 1990; Hornik 1991) established that three layers are sufficient for approximations of general continuous functions.

Nevertheless, the approach based on the technique developed by Kolmogorov is not without value. The above mentioned theorems are proved very elegantly using advanced theorems from functional analysis. However, nondirect proofs do not provide clear insight into constructions of approximating functions. The directness of our proofs can

be exploited for estimating the number of hidden units and for exploring which properties of a function being approximated are relevant for the growth of this number. The first step in this direction was done in Kůrková (1991), where the numbers of units in the second and the third layer are estimated by $nm(m+1)$ and $m^2(m+1)^n$, respectively, where n is the dimension of the unit cube I^n and m depends on $\epsilon/\|f\|$ as well as on the rate with which f increases distances. Hopefully, further analysis could bring finer estimates and more insight to the questions of what properties of the function being implemented play a role in determining the number of hidden units, and whether this number can be sufficiently reduced by using two instead of only one hidden layer.

References

Alexandrov, P. S. (ed.) 1983. *Die Hilbertschen Probleme.* Akademische Verlagsgesellschaft, Leipzig.

Arnold, V. I. 1957. On functions of three variables. *Dokl. Akad. Nauk USSR* **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, 607–611. IEEE, New York.

Cybenko, G. 1989. Approximation by superpositions of a single function. *Math. Control, Signals Syst.* **2**, 303–314.

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

Girosi, F., and Poggio, T. 1989. Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Comp.* **1**, 465–469.

Hecht-Nielsen, R. 1987. Kolmogorov's mapping neural network existence theorem. In *Proceedings of the International Conference on Neural Networks*, pp. III, 11–14. IEEE, New York.

Hecht-Nielsen, R. 1989. Theory of the back-propagation neural network. In *Proceedings of the International Joint Conference on Neural Networks*, pp. I, 593–606. IEEE, New York.

Hecht-Nielsen, R. 1990. *Neurocomputing.* Addison-Wesley, New York.

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

Hornik, K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **2**, 251–257.

Kolmogorov, A. N. 1957. On the representations of continuous functions of many variables by superpositions of continuous functions of one variable and addition. *Dokl. Akad. Nauk USSR* **114** (5), 953–956.

Kůrková, V. 1991. Kolmogorov's theorem and multilayer neural networks. *Neural Networks* (in press).

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

Mandelbrot, B. B. 1982. *The Fractal Geometry of Nature.* Freeman, San Francisco.

Sprecher, D. A. 1965. On the structure of continuous functions of several variables. *Trans. Am. Math. Soc.* **115**, 340–355.

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, 613–617. IEEE, New York.

Stinchcombe, M., and White, H. 1990. Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights. In *Proceedings of the International Joint Conference on Neural Networks*, pp. III, 7–16. IEEE, New York.

Received 20 December 1991; accepted 6 June 1991.

Communicated by Rodney Goodman

An Exponential Response Neural Net

Shlomo Geva
Joaquin Sitte
Faculty of Information Technology,
Queensland University of Technology, GPO Box 2434, Brisbane Q 4001, Australia

By using artificial neurons with exponential transfer functions one can design perfect autoassociative and heteroassociative memory networks, with virtually unlimited storage capacity, for real or binary valued input and output. The autoassociative network has two layers: input and memory, with feedback between the two. The exponential response neurons are in the memory layer. By adding an encoding layer of conventional neurons the network becomes a heteroassociator and classifier. Because for real valued input vectors the dot-product with the weight vector is no longer a measure for similarity, we also consider a euclidean distance based neuron excitation and present Lyapunov functions for both cases. The network has energy minima corresponding only to stored prototype vectors. The exponential neurons make it simpler to build fast adaptive learning directly into classification networks that map real valued input to any class structure at its output.

1 Introduction

Associative memories implemented on neural networks have been the subject of intensive research for some years now. The associative neural network memories tend to have low storage capacity (Gardner 1987), generate spurious states, or use global learning rules. For polar, binary inputs many of these problems can be overcome, with the use of exponential response neurons (Goodman and Chiueh 1988), to control the competition and cooperation between the prototype vectors (Geva and Sitte 1990a). Chiueh and Goodman (1988) first proposed a binary associative memory with exponential neurons and showed that the storage capacity for random memory vectors grows exponentially with the dimesion of the vectors. They also described a VLSI implementation of their network (Goodman and Chiueh 1990; Chiueh and Goodman 1991). Exponential response neurons have also been used by Specht (1990) in his probabilistic neural network (PNN). The purpose of this paper is to describe an exponential response network for real valued inputs, to analyze its dynamics through a Lyapunov function, and to show how it can be used for classification, with fast learning built into the network. Most of

