LA-UR- 98-3462
Approved for public release; distribution is unlimited.

Title: LARGER BASES AND MIXED ANALOG/DIGITAL NEURAL NETS

CONF-981129--

Author(s): Valeriu C. Beiu, NIS-1

Submitted to: Artificial Neural Networks in Engineering '98 St. Louis, 1-4 November, 1998

Los Alamos
NATIONAL LABORATORY

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

LARGER BASES AND MIXED ANALOG/DIGITAL NEURAL NETS

VALERIU BEIU [1]
Los Alamos National Laboratory, Division NIS-1, MS D466
Los Alamos, New Mexico 87545, USA (e-mail: beiu@lanl.gov*)*

ABSTRACT:
The paper overviews results dealing with the approximation capabilities of neural networks, and bounds on the *size* of threshold gate circuits. Based on an explicit numerical algorithm for Kolmogorov's superpositions we will show that minimum *size* neural networks—for implementing any Boolean function—have the identity function as the activation function. Conclusions and several comments on the required precision are ending the paper.

INTRODUCTION

In this paper a *network* is an acyclic graph having several input nodes, and some (at least one) output nodes. If a synaptic *weight* is associated with each edge, and each node computes the weighted sum of its inputs to which a nonlinear activation function is then applied: $f(x_1, \ldots, x_\Delta) = \sigma(\sum_{i=1}^{\Delta} w_i x_i + \theta)$, the network is a *neural network* (NN), with w_i the synaptic *weights*, θ known as the *threshold*, Δ being the *fan-in*, and σ a non-linear activation function. NNs are commonly characterised by: *depth* (*i.e.*, number of layers), and *size* (*i.e.*, number of neurons).

The paper starts by overviewing results dealing with the approximation capabilities of NNs, and details upper and lower bounds showing that arbitrary *Boolean functions* (BFs) require exponential *size threshold gate* (TG) *circuits* (TGCs). Based on a constructive solution for Kolmogorov's superpositions, *size*-optimal NNs for implementing any BF, *the nonlinear activation function of the neurons is the identity function.* Because both Boolean and TGCs require exponential *size*, it follows that *size*-optimal implementations of BFs can be obtained only in analog circuitry. Conclusions, and several comments on the required precision are ending the paper.

PREVIOUS RESULTS

NNs have been experimentally shown to be quite effective in many applications (see *Applications of Neural Networks* in (Arbib, 1995), and *Part F* and *G* from (Fiesler & Beale, 1996)). This success has led researchers to undertake a rigorous analysis of their mathematical properties and has generated two directions of research for finding: (*i*) existence/constructive proofs for the '*universal approximation problem*'; (*ii*) tight bounds on the *size* needed by the approximation problem. The paper will focus on both aspects, for the case when the functions to be implemented are BFs.

[1] On leave of absence from the "Politehnica" University of Bucharest, Computer Science Department, Spl. Independenței 313, RO-77206 Bucharest, România.

Neural Networks as Universal Approximators

The first line of research has concentrated on the approximation capabilities of NNs (Blum & Li, 1991; Ito, 1991). It was started by Hecht-Nielsen (1987), Lippmann (1987), and LeCun (1987), who were probably the first to recognise that the specific format in (Sprecher, 1965) $f(x_1, ..., x_n) = \sum_{q=1}^{2n+1} \{\Phi_q [\sum_{p=1}^{n} \alpha_p \psi (x_p + qa)]\}$ of Kolmogorov's superpositions $f(x_1, ..., x_n) = \sum_{q=1}^{2n+1} \Phi_q (y_q)$ (Kolmogorov, 1957) can be interpreted as a NN with one hidden layer. This gave an existence proof of the approximation properties of NNs. The first nonconstructive proof was given by Cybenko (1989) using a continuous activation function, and was independently presented by Irie and Miyake (1988). Thus, the fact that NNs are computationally universal when modifiable connections are allowed, was established. Different enhancements have been later presented (for details see (Beiu, 1998c; Scarselli & Tsoi, 1998)):

- Funahashi (1989) proved the same result in a more constructive way and refined the use of Kolmogorov's theorem in (Hecht-Nielsen, 1987), giving an approximation result for two-hidden-layer NNs;
- Hornik *et al.* (1989) showed that the continuity requirement for the output function can partly be removed, and later (Hornik *et al.*, 1990) proved that a NN can approximate simultaneously a function and its derivative;
- Park and Sandberg (1991) used radial basis functions in the hidden layer, and gave an almost constructive proof;
- Hornik (1991) showed that the continuity requirement can be completely removed, the activation function having to be 'bounded and nonconstant';
- Geva and Sitte (1992) proved that four-layered NNs with sigmoid activation function are universal approximators;
- Kůrková (1992) has demonstrated the existence of approximate superposition representations within the constraints of NNs, *i.e.* ψ and Φ_q can be approximated with functions of the form $\sum a_r \sigma (b_r x + c_r)$;
- Mhaskar and Micchelli (1992) approach was based on truncating the infinite sum of the Fourier series of the function to a finite set, and rewriting e^{ikx} in terms of the activation function;
- Koiran (1993) presented a more general proof than (Funahashi, 1989), as it allows the use of units with 'piecewise continuous' activation functions;
- Leshno *et al.* (1993) relaxed the condition for the activation function to 'locally bounded piecewise continuous';
- Hornik (1993) added to these results by proving that: (*i*) if the activation function is locally Riemann integrable and nonpolynomial, the *weights* and the *thresholds* can be constrained to arbitrarily small sets; and (*ii*) if the activation function is locally analytic, a single universal *threshold* will do;
- Barron (1993) described spaces of functions that can be approximated by the relaxed algorithm (Jones, 1992) using functions computed by single-hidden-layer networks of perceptrons;
- Attali and Pagès (1997) provided an elementary proof based on the Taylor expansion and the Vandermonde determinant, yielding bounds for the design of the hidden layer and convergence results for the derivatives.

All these results—except partly (Park & Sandberg, 1991; Kůrková, 1992; Barron, 1993; Koiran, 1993))—were obtained provided that sufficiently many hidden units are available. More constructive solutions have been obtained in small *depth* (Nees, 1994), but their *size* still grows fast with respect to the number of dimensions and/or examples, or with the required precision. Recently, an explicit numerical algorithm for superpositions has been detailed (Sprecher, 1996, 1997).

Threshold Gate Circuits

The other line of research was to find the smallest *size* NN which can realise an arbitrary function given a set of m vectors from $\mathbb{R}^n$. Many results have been obtained for TGs (Minnik, 1961). The first lower bound on the *size* of a TGC for almost all n-ary BFs was $size \geq 2\,(2^n/n)^{1/2}$ (Neciporuk, 1964). Later a very tight upper bound was proven in $depth = 4$: $size \leq 2\,(2^n/n)^{1/2} \times \{1 + \Omega\,[(2^n/n)^{1/2}]\}$ (Lupanov, 1973).

For classification problems, it was known that a NN of $depth = 3$ and $size = m - 1$ could compute an arbitrary dichotomy. The main improvements have been: (*i*) Baum (1988) presented a TGC with one hidden layer having $\lceil m/n \rceil$ neurons for a set of m points *in general position* in $\mathbb{R}^n$; if the points are on the corners of the n-dimensional hypercube, $m - 1$ nodes are still needed; (*ii*) a slightly tighter bound of only $\lceil 1 + (m-2)/n \rceil$ neurons in the hidden layer for a set of m points which satisfy a more relaxed topological assumption was proven in (Huang & Huang, 1991); the $m - 1$ nodes condition was shown to be the least upper bound needed; (*iii*) Arai (1993) showed that $m - 1$ hidden neurons are necessary for arbitrary separability, but improved the bound for the dichotomy problem to $m/3$; (*iv*) Beiu and De Pauw (1997) have detailed lower and upper bounds by estimating the entropy of the data-set $2m/(n\log n) < size < 1.44m/n$ (see (Beiu et al., 1998)).

Bulsari (1993) has tried to unify these two lines of research by first presenting analytical solutions for the general NN problem in one dimension (having infinite *size*), and then giving practical solutions for the one-dimensional cases (*i.e.*, upper bounding the *size*). Extensions to the n-dimensional case using three- and four-layers solutions were derived under piecewise constant and piecewise linear approximations.

Boolean Functions

The particular case of BFs has been studied intensively (Parberry, 1994; Beiu, 1998c). Many results have been obtained for particular BFs (Siu *et al.*, 1991), but a *size*-optimal result for BFs that have exactly m groups of ones in their truth table $\mathbb{F}_{n,m}$ was detailed by Red'kin (1970): *"The complexity realisation (i.e., number of threshold elements) of* $\mathbb{F}_{n,m}$ *is at most* $2\,(2m)^{1/2} + 3$*."* This result is valid for unlimited *fan-in* TGs. Departing from these lines, Horne and Hush (1994) have detailed a solution for limited *fan-in* TGCs: *"Arbitrary BFs of the form* $f : \{0,1\}^n \rightarrow \{0,1\}^m$ *can be implemented in a NN of perceptrons restricted to fan-in 2 with a node complexity of* $\Theta\,\{m\,2^n/(n + \log m)\}$ *and requiring* $\boldsymbol{O}\,(n)$ *layers."*

SIZE-OPTIMAL IMPLEMENTATIONS

It is known that arbitrary BFs require exponential *size* AND-OR circuits. As has been seen, the known bounds for *size* are also exponential if TGCs are used. These bounds reveal exponential gaps, and suggest that TGCs with more layers might have a smaller *size* (Beiu, 1998b; Beiu & Makaruk, 1998).

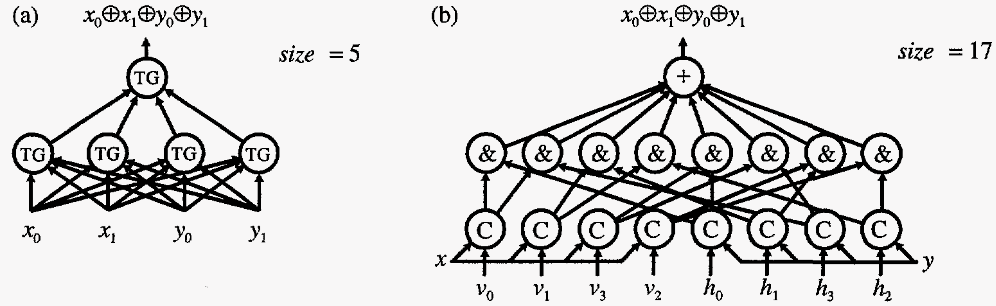

Figure 1: The PARITY problem: (a) TG solution; (b) approximating a 4-input BF (see (Beiu, 1998a)).

A different approach is to use Kolmogorov's superpositions, which shows that there are NNs having only $2n + 1$ (*size*-optimal) neurons which can approximate any function. We start from (Sprecher, 1996, 1997): *"Define the function* $\psi : \mathcal{E} \rightarrow \mathcal{D}$ *such that for each integer* $k \in N$: $\psi\ (\sum_{r=1}^{k} i_r \gamma^{-r}) = \sum_{r=1}^{k} \tilde{i}_r 2^{-m_r} \gamma^{-(n^{r-m_r}-1)/(n-1)}$ *where* $\tilde{i}_r = i_r - (\gamma - 2)\langle i_r \rangle$ *and* $m_r = \langle i_r \rangle \{1 + \sum_{s=1}^{r-1} [i_s] \times \ldots \times [i_{r-1}]\}$ *for* $r = 1, 2, \ldots, k$." Here $\gamma \geq 2n + 2$ is a base, $\mathcal{E} = [0, 1]$ is the unit interval, $\mathcal{D}$ is the set of terminating rational numbers $d_k = \sum_{r=1}^{k} i_r \gamma^{-r}$ defined on $k \in N$ digits ($0 \leq i_r \leq \gamma - 1$). Also, $\langle i_1 \rangle = [i_1] = 0$, while for $r \geq 2$: $\langle i_r \rangle = 0$ when $i_r = 0, 1, \ldots, \gamma - 2$, $\langle i_r \rangle = 1$ when $i_r = \gamma - 1$, $[i_r] = 0$ when $i_r = 0, 1, \ldots, \gamma - 3$, while $[i_r] = 1$ when $i_r = \gamma - 2, \gamma - 1$.

If the functions to be 'approximated' are BFs, one digit is enough $\psi\ (0.i_1) = 0.i_1$, *i.e.*, the identity function $\psi\ (x) = x$. Such a solution builds simple analog neurons. They have *fan-in* $\Delta \leq 2n + 1$, for which the known *weight* bounds are (Myhill & Kautz, 1961; Parberry, 1994): $2^{(\Delta-1)/2} < weight < (\Delta + 1)^{(\Delta+1)/2} / 2^{\Delta}$ ($\Delta \geq 4$). Thus, a *precision* of between Δ, and $\Delta \log\Delta$ bits per *weight* would be expected. Unfortunately, the constructive solution for Kolmogorov's superpositions requires a double exponential precision for ψ, and $\alpha_p = \sum_{r=1}^{\infty} \gamma^{-(p-1)(n^r-1)/(n-1)}$. For BFs, this precision becomes $(2n + 2)^{-n}$, or $2n\log n$ bits per *weight*, while analog implementations are limited to just several bits of precision (Kramer, 1996). A possible solution are algorithms relying on limited integer *weights* (Drāghici & Sethi, 1997; Beiu, 1998a).

Let us consider the PARITY function of four bits. It is known that PARITY can be implemented with three 2-input XOR gates, or with five 4-input TGs (Fig. 1(a)). In general, a 4-input BF requires $2\sqrt{16} + 3 = 11$ TGs (Red'kin, 1970). A classical Boolean solution requires eight AND gates. Another solution would approximate a 4-input BF (Fig. 1(b)). We consider two analog inputs: $\overline{x_0\, x_1}$, and $\overline{y_0\, y_1}$. Fig. 2(a) presents the modified Karnaugh map, and Fig. 2(b) the solution. It has ψ = COMPARISON, and the inputs are "translated" with fixed constants. The $2n + 1 = 5$ hidden functions are AND functions (Φ = AND), while the addition is an OR function.

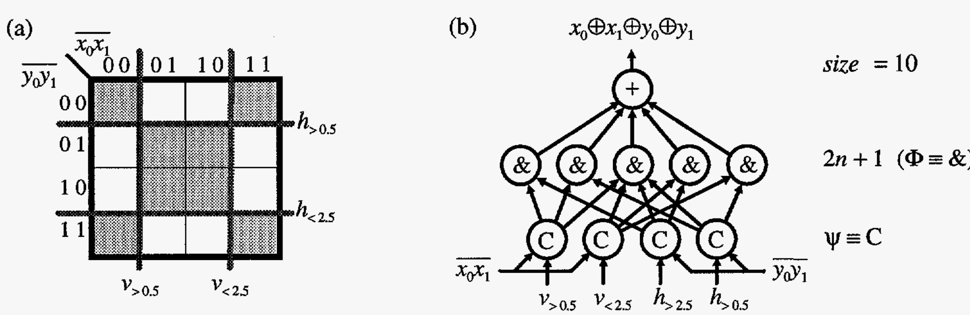

Figure 2: The PARITY problem: (a) modified Karnaugh map; (b) solution using COMPARISONS (C).

Due to the limited precision, an optimal should decompose the complex BFs in simpler BFs which can be implemented using Kolmogorov's superpositions. The partial results from this first layer can be combined again using Kolmogorov's superpositions. The final implementation will requires more than three layers. Thus, a systematic solution which would utilise silicon to the best advantage would be to rewrite a given computation in a base larger than 2, and use Kolmogorov's superpositions for the analog implementation of the digit-wise computations in this larger base.

CONCLUSIONS

Arbitrary BFs can be implemented using: (*i*) classical Boolean gates in exponential *size*; (*ii*) TGs in exponential *size* (still, there are exponential gaps between classical Boolean solutions and TG ones); (*iii*) analog building blocks in linear *size* (having linear *fan-in* and polynomial precision *weights* and *thresholds*), the nonlinear activation function being the identity function. It follows that *size*-optimal hardware implementations of BFs can be obtained only using analog (or mixed analog/digital) circuitry. The high precision required by Kolmogorov's superpositions can be tackled by decomposing a complex BF into simpler BFs (equivalent to computing in a larger base). Due to the reduced number of inputs, Kolmogorov's superpositions can be used to design the analog implementations of the digit-wise computations in this larger base.

REFERENCES

M. Arai. Bounds on the Number of Hidden Units in Binary-valued Three-layer Neural Networks. *Neural Networks*, **6**, 855-60, 1993.

M.A. Arbib (ed.). *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, 1995.

J.-G. Attali & G. Pagès. Approximations of Functions by a Multilayer Perceptron: a New Approach. *Neural Networks*, **10**, 1069-81, 1997.

A.R. Barron. Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Trans. Info. Th.*, **39**, 930-45, 1993.

E.B. Baum. On the Capabilities of Multilayer Perceptrons. *J. Complexity*, **4**, 193-215, 1988.

V. Beiu. Reduced Complexity Constructive Learning Algorithm. *Neural Network World*, **8**, 1-38, 1998a.

V. Beiu. On the Circuit and VLSI Complexity of Threshold Gate COMPARISON. *Neurocomp.*, **19**, 77-98, 1998b.

V. Beiu. *VLSI Complexity of Discrete Neural Networks*. Gordon & Breach, Newark, NJ, 1998c.

V. Beiu & T. De Pauw. Tight Bounds on the Size of Neural Networks for Classification Problems. In J. Mira *et al.* (eds.): *Biological and Artificial Computation*, Springer, Berlin, 743-52, 1997.

V. Beiu & H.E. Makaruk. Deeper Sparser Nets Can Be Optimal. *Neural Proc. Lett.*, **8**, 1998.

V. Beiu *et al.* A Constructive Approach to Calculating Lower Entropy Bounds. *Neural Proc. Lett.*, **8**, 1998.

E. Blum & K. Li. Approximation Theory and Feedforward Networks. *Neural Networks*, **4**, 511-5, 1991.

A. Bulsari. Some Analytical Solutions to the General Approximation Problem for Feedforward Neural Networks. *Neural Networks*, **6**, 991-6, 1993.

G. Cybenko. Approximations by Superpositions of a Sigmoid Function. *Math. of Control, Signals and Systems*, **2**, 303-14, 1989.

S. Drăghici & I.K. Sethi. On the Possibilities of the Limited Precision Weights Neural Networks in Classification Problems. In J. Mira *et al.* (eds.): *Biological and Artificial Computation*, Springer, Berlin, 753-62, 1997.

E. Fiesler & R. Beale (eds.). *Handbook of Neural Computation*. Inst. of Physics, New York, 1996.

K.-I. Funahashi. On the Approximate Realization of Continuous Mapping by Neural Networks. *Neural Networks*, **2**, 183-92, 1989.

S. Geva & J. Sitte. A Constructive Method for Multivariate Function Approximation by Multilayered Perceptrons. *IEEE Trans. Neural Networks*, **3**, 621-3, 1992.

R. Hecht-Nielsen. Kolmogorov's Mapping Neural Network Existence Theorem. *Proc. Intl. Conf. on Neural Networks*, IEEE CS Press, CA, 11-4, 1987.

B.G. Horne & D.R. Hush. On the Node Complexity of Neural Networks. *Neural Networks*, **7**, 1413-26, 1994.

K. Hornik. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Network*, **4**, 251-7, 1991.

K. Hornik. Some New Results on Neural Network Approximation. *Neural Network*, **6**, 1069-72, 1993.

K. Hornik *et al.* Multilayer Feedforward Neural Networks Are Universal Approximators. *Neural Network*, **2**, 359-66, 1989.

K. Hornik *et al.* Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks. *Neural Network*, **3**, 551-60, 1990.

S.-C. Huang & Y.-F. Huang. Bounds on the Number of Hidden Neurons of Multilayer Perceptrons in Classification and Recognition. *IEEE Trans. Neural Networks*, **2**, 47-55, 1991.

B. Irie & S. Miyake. Capabilities of Three-Layered Perceptrons. *Proc. Intl. Conf. on Neural Networks*, IEEE CS Press, CA, vol. **1**, 641-8, 1988.

Y. Ito. Approximation of Functions on a Compact Set by Finite Sums of Sigmoid Functions without Scaling. *Neural Networks*, **4**, 817-26, 1991.

L.K. Jones. A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. *Ann. Statist.*, **20**, 608-13, 1992.

P. Koiran. On the Complexity of Approximating Mappings Using Feedforward Networks. *Neural Networks*, **6**, 649-53, 1993.

A.N. Kolmogorov. On the Representation of Continuous Functions of Many Variables by Superposition of Continuous Functions of One Variable and Addition. *Dokl. Akad. Nauk SSSR*, **114**, 953-6, 1957.

A.H. Kramer. Array-Based Analog Computation; Principles, Advantages and Limitations. *Proc. Microelectronics for Neural Networks*, IEEE CS Press, CA, 68-79, 1996.

V. Kůrková. Kolmogorov's Theorem and Multilayer Neural Networks. *Neural Networks*, **5**, 501-6, 1992.

Y. LeCun. Models connexionistes de l'apprentisage. *MSc Th.*, Univ. Pierre et Marie Curie, Paris, 1987.

M. Leshno, V.Y. Lin, A. Pinkus & S. Schocken. Multilayer Feedforward Neural Networks with a Non-polynomial Activation Function Can Approximate any Function. *Neural Networks*, **6**, 861-7, 1993.

R.P. Lippmann. An Introduction to Computing with Neural Nets. *IEEE ASSP Mag.*, **4**, 4-22, 1987.

O.B. Lupanov. The Synthesis of Circuits from Threshold Elements. *Problemy Kibernetiki*, **20**, 109-40, 1973.

H.N. Mhaskar & C. Micchelli. Approximation by Superposition of Sigmoidal and Radial Basis Functions. *Advs. in Appl. Maths.*, **13**, 350-73, 1992.

R.C. Minnik. Linear-Input Logic. *IRE Trans. Electr. Comp.*, **10**, 6-16, 1961.

J. Myhill & W.H. Kautz. On the Size of Weights Required for Linear-Input Switching Functions. *IRE Trans. Electr. Comp.*, **10**, 288-90, 1961.

E.I. Neciporuk. The Synthesis of Networks from Threshold Elements. *Soviet Mathematics–Doklady*, **5**, 163-6, 1964.

M. Nees. Approximate Versions of Kolmogorov's Superposition Theorem, Proved Constructively. *J. Comp. Appl. Math.*, **54**, 239-50, 1994.

I. Parberry. *Circuit Complexity and Neural Networks*. MIT Press, Cambridge, 1994.

J. Park & I.W. Sandberg. Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, **3**, 246-57, 1991.

N.P. Red'kin. Synthesis of Threshold Circuits for Certain Classes of Boolean Functions. *Kibernetika*, **5**, 6-9, 1970.

F. Scarselli & A.C. Tsoi. Universal Approximation Using Feedforward Neural Networks: A Survey of Some Existing Methods, and Some New Results. *Neural Networks*, **11**, 15-37, 1998.

K.-Y. Siu *et al.* Depth-Size Tradeoffs for Neural Computations. *IEEE Trans. Comp.*, **40**, 1402-12, 1991.

D.A. Sprecher. On the Structure of Continuous Functions of Several Variables. *Trans. American Math. Soc.*, **115**, 340-55, 1965.

D.A. Sprecher. A Numerical Implementation of Kolmogorov's Superpositions. *Neural Networks*, **9**, 765-72, 1996.

D.A. Sprecher. A Numerical Implementation of Kolmogorov's Superpositions II. *Neural Networks*, **10**, 447-57, 1997.