Title: ON KOLMOGOROV'S SUPERPOSITIONS AND BOOLEAN FUNCTIONS

CONF-981210 --

Author(s): Valeriu Beiu, NIS-1

Submitted to: Brazilian Symposium on Neural Networks
Belo Horizonte, Brazil, Dec. 9-11, 1998

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. The Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
On Kolmogorov’s Superpositions and Boolean Functions

Valeriu Beiu*

Space and Atmospheric Division NIS-1, MS D466
Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
E-mail: beiu@lanl.gov

Abstract

The paper overviews results dealing with the approximation capabilities of neural networks, as well as bounds on the size of threshold gate circuits. Based on an explicit numerical (i.e., constructive) algorithm for Kolmogorov’s superpositions we will prove that in order to obtain linear size NNs (i.e., size-optimal) for implementing any Boolean function, the nonlinear activation function of the neurons has to be the identity function. Because classical AND-OR implementations, as well as threshold gate implementations require exponential size, it follows that size-optimal implementations of discrete NNs (i.e., implementing BFs) can be obtained only in analog circuitry. Conclusions, and several comments on the required precision are ending the paper.

Keywords: neural networks, Kolmogorov’s superpositions, Boolean circuits, threshold gate circuits, analog circuits, size, precision.

1 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 non-linear activation function is then applied:

\[ f(x_1, \ldots, x_n) = \sum w_i x_i + \theta \]

the network is a neural network (NN), with \( w_i \in \mathbb{R} \) the synaptic weights, \( \theta \in \mathbb{R} \) known as the threshold, and \( \sigma \) a non-linear activation function. Because the underlying graph is acyclic, the network does not have feedback, and can be layered. That is why such a network is also known as a multilayer feedforward neural network, and is commonly characterised by: its depth (i.e., number of layers), and its size (i.e., number of neurons).

The paper starts by overviewing several of the results dealing with the approximation capabilities of NNs. They are followed by details on known upper and lower bounds on the size of threshold gate circuits (TGCs), which show that TGCs require exponential size for implementing arbitrary Boolean functions (BFs). Based on a constructive solution for Kolmogorov’s superpositions we will prove that in order to obtain linear size NNs (i.e., size-optimal) for implementing any BF, the nonlinear activation function of the neurons has to be the identity function. Because both Boolean circuits and TGCs require exponential size, it follows that size-optimal implementations of discrete NNs (i.e., implementing BFs) can be obtained only in analog circuitry. Conclusions, and several comments on the required precision are ending the paper.

2 Previous Results

NNs have been experimentally shown to be quite effective in many applications (see Applications of Neural Networks in [2], together with Part F: Applications of Neural Computation and Part G: Neural Networks in Practice: Case Studies from [20]). 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.

2.1 Neural Networks as Universal Approximators

The first line of research has concentrated on the approximation capabilities of NNs [14, 22, 35, 36]. It was started in 1987 by Hecht-Nielsen [26] and Lippmann [47] who, together with LeCun [45], were probably the first to recognise that the specific format in [67, 68] of the form:

\[ f(x_1, \ldots, x_n) = \sum_{q=1}^{2^a+1} \Phi_q \left[ \sum_{p=1}^{n} \alpha_p \psi(x_p + q\theta) \right] \] (1)

of Kolmogorov’s superpositions [41]: \( f(x_1, \ldots, x_n) = \sum_{q=1}^{2^a+1} \Phi_q (x_q) \) 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 in 1988 by Cybenko [16, 17] using a continuous activation function, and was independently presented by Irie
and Miyake [34]. Similar results for radial basis functions were shortly reported [24, 59]. Thus, the fact that NNs are computationally universal—with more or less restrictive conditions—when modifiable connections are allowed, was established. Different enhancements have been later presented (for details see [64], and Chp. 1 in [9]):

- Funahashi [21] proved the same result but in a more constructive way and also refined the use of Kolmogorov’s theorem in [26], giving an approximation result for two-hidden-layer NNs;
- Hornik et al. [31] showed that the continuity requirement for the output function can partly be removed;
- Hornik et al. [32] also proved that a NN can approximate simultaneously a function and its derivative;
- Park and Sandberg [57, 58] used radial basis functions in the hidden layer, and gave an almost constructive proof;
- Hornik [29] showed that the continuity requirement can be completely removed, the activation function having to be “bounded and nonconstant”;
- Geva and Sitte [23] proved that four-layered NNs with sigmoid activation function are universal approximators;
- Kůrková [43, 44] has demonstrated the existence of approximate superposition representations within the constraints of NNs, i.e. ψ and Φ can be approximated with functions of the form \( \sum a_i \sigma (b_i + x + c_i) \), where \( \sigma \) is an arbitrary activation sigmoidal function;
- Mhaskar and Michelli [49, 50] approach was based on the Fourier series of the function, by truncating the infinite sum to a finite set, and rewriting \( e^{ux} \) in terms of the activation function (which now has to be periodic);
- Koiran [40] presented a new proof on the line of Funahashi’s proof [21], but more general in that it allows the use of units with “piecewise continuous” activation functions; these include the particular but important case of threshold gates (TGs);
- Leshno et al. [46] relaxed the condition for the activation function to “locally bounded piecewise continuous” (i.e., if and only if the activation function is not a polynomial), thus embedding as special cases almost all the activation functions that have been reported in the literature;
- Hornik [30] 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;
- Funahashi and Nakamura [22] showed that the universal approximation theorem also holds for trajectories of patterns;
- Sprecher [69] has demonstrated that there are universal hidden layers that are independent of the number of input variables \( n \);
- Barron [4] described spaces of functions that can be approximated by the relaxed algorithm of Jones [37] using functions computed by single-hidden-layer networks of perceptrons;
- Attali and Pagès [3] provide 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—with the partial exception of [4, 40, 43, 57, 58]—were obtained “provided that sufficiently many hidden units are available” (i.e., with no claims on the size minimality). More constructive solutions have been obtained in very small depth later [38, 54, 55], 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 [70–72].

### 2.2 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 [51]. The first lower bound of:

\[
\text{size} \geq 2 (2^{n/n})^{1/2}
\]

on the size of a TGC for “almost all” \( n \)-ary BFs (i.e., \( f : \mathbb{B}^n \rightarrow \mathbb{B} \)) was given in [53]. Later [48] a very tight upper bound was proven in depth = 4:

\[
\text{size} \leq 2 (2^{n/n})^{1/2} \times \{1 + \Omega ([2^{n/n})^{1/2}]\}
\]

A similar existence exponential lower bound of \( \Omega (2^{n/3}) \) for arbitrary BFs can be found in [65], which also gives bounds for many particular but important BFs (see also [63]).

For classification problems \( (f : \mathbb{R}^n \rightarrow \mathbb{B}^k) \), the first result was that a NN of depth = 3 and size = \( m - 1 \) could compute an arbitrary dichotomy. The main improvements have been:

- Baum [5] presented a TGC with one hidden layer having \([m/n])\) neurons capable of realising an arbitrary dichotomy on a set of \( m \) points in general position in \( \mathbb{R}^n \); if the points are on the corners of the \( n \)-dimensional hypercube \( (f : \mathbb{B}^n \rightarrow \mathbb{B}) \), \( m - 1 \) nodes are still needed;
- a slightly tighter bound of only \([1 + (m - 2)/n])\) neurons in the hidden layer for realising an arbitrary dichotomy on a set of \( m \) points which satisfy a more relaxed topological assumption was proven in [33];
the $m-1$ nodes condition was shown to be the least upper bound needed;
- Arai [1] showed that $m-1$ hidden neurons are necessary for arbitrary separability, but improved the bound for the dichotomy problem to $m/3$ (without any condition on the inputs);
- Beiu and De Pauw [10] have detailed existence lower and upper bounds: $2m/(n\log n) < \text{size} < 1.44m/n$ by estimating the entropy of the data-set ([11, 13]).

Other existence lower bounds for the arbitrary dichotomy problem can be found in [25, 59]:
- a depth-2 TGC requires at least $m/\{n \log(m/n)\}$ TGs;
- a depth-3 TGC requires at least $2\{m/\log m\}^{1/2}$ TGs in each of the two hidden layer (if $m \gg n^2$);
- an arbitrarily interconnected TGC without feedback needs $2\{m/\log m\}^{1/2}$ TGs (if $m \gg n^2$).

One study [15] 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., including an upper bound on the size). Extensions to the $n$-dimensional case using three- and four-layers solutions were derived under piecewise constant approximations (having constant or variable width partitions), and under piecewise linear approximations (using ramps instead of sigmoids).

### 2.3 Boolean Functions

The particular case of BFs has been studied intensively [9, 65]. Many results have been obtained for particular BFs [63, 65], but a size-optimal result for BFs that have exactly $m$ groups of ones in their truth table $F_{n,m}$ (BFs which are defined on the $m$ groups) was detailed by Red’kin in 1970 [62].

**Theorem from [62]** The complexity realization (i.e., number of elements) of $F_{n,m}$ is at most $2(2m)^{1/2} + 3$.

This result—as all the previous mentioned ones—is valid for unlimited fan-in TGs. Departing from these lines, Horne and Hush [28] have detailed a solution for limited fan-in TGCs.

**Theorem from [28]** Arbitrary Boolean functions 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 $\Theta \{m \cdot 2^n / (n + \log m)\}$ and requiring $O(n)$ layers.

### 3 Size-optimal Implementations

It is known that implementing arbitrary BFs using classical Boolean gates (i.e., AND and OR gates) requires exponential size circuits. As has been seen from all the previous results, the known bounds for size are also exponential if TGCs are used for solving arbitrary BFs [6]. It is true that these bounds reveal exponential gaps (thus encouraging research efforts to reduce them), and also suggest that TGCs with more layers (depth $\neq$ small constant [8, 12]) might have a smaller size.

A different approach is to use Kolmogorov’s superpositions, which shows that there are NNs having only $2n + 1$ neurons which can approximate any function. Such a solution would clearly be size-optimal. We start from [70–72], where a constructive solution for the general case was detailed.

**Theorem from [70]** Define the function $\psi: \delta \rightarrow \mathcal{D}$ such that for each integer $k \in N$:

$$\psi \left( \sum_{r=1}^{k} i, \gamma^{-r} \right) = \sum_{r=1}^{k} \bar{i}, 2^{-m_r} \gamma \frac{n-r-m_{r-1}}{n-1}$$

where $\bar{i}_r = i_r - (\gamma-1) \langle i_r \rangle$ 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, $\delta = \{0, 1\}$ is the unit interval, $\mathcal{D}$ is the set of 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, $\langle i_r \rangle = [i_r] = 0$, 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 we limit the functions to be ‘approximated’ to BFs, one digit is enough ($k = 1$), which gives $\psi(0, i) = 0, i$, which is the identity function $\psi(x) = x$. Such a solution builds quite simple analog neurons. They have fan-in $\Delta \leq 2n + 1$, for which the known weight bounds (holding for any fan-in $\Delta \geq 4$) are [52, 56, 61, 66]:

$$2^{(\Delta-1)/2} < \text{weight} < (\Delta + 1)^{(\Delta+1)/2} / 2^{\Delta}$$

Thus, a precision of between $\Delta$, and $\Delta \log \Delta$ bits per weight would be expected. Unfortunately, the constructive solution for Kolmogorov’s superpositions requires a double exponential precision for $\psi$ (eq. 4), and for the weights:

$$\alpha_{p} = \sum_{r=1}^{\infty} \gamma^{-(p-1)r} \frac{n-r}{n-1}.$$  

For BFs precision is reduced to $2n^{+2}$ or $2n \log n$ bits per weight. Analog implementations are limited to just several bits of precision [42], this being one of the reasons for investigations on precision [18, 27, 73, 74] and on algorithms relying on limited integer weights [7, 19, 39].

As an example, let us consider the PARITY function of four bits. It is known that PARITY can be implemented with just three 2-input XOR gates (Fig. 1.a), or with five 4-input TGs (Fig. 1.b). It is also known that, in general, a 4-input
BF requires $2\sqrt{16} + 3 = 11$ TGs [62]. A classical Boolean solution requires eight (one for each minterm) AND gates (Fig. 1.c). A brute force solution would approximate a 4-input BF (see Fig. 2.a and 2.b). A hand crafted solution for the same case can be seen in Fig. 2.c. The next step is to consider $X=x_0x_1$, and $Y=y_0y_1$. In this particular case, a solution is presented in Fig. 3.a (a hand crafted solution can be seen in Fig. 3.b). They have COMPARISON as the $v$ function, and the inputs are “translated” with fixed constants. The $2n+1=5$ hidden functions ($\Phi_5$) are simple AND functions, while—departing from eq. (1)—the addition is replaced by an OR function.

Due to the limitation on precision an optimal solution for implementing BFs should decompose the given BF in simpler BFs which can be efficiently implemented based on Kolmogorov’s superpositions (i.e., we have to reduce $n$ to small values). The partial results from this first layer of analog building blocks can be combined using again Kolmogorov’s superpositions. The final analog implementation will require more than three layers. It follows that a systematic solution which would utilise silicon to the best advantage would be to rewrite a given computation (i.e., set of BFs) in a base larger than 2, and use Kolmogorov’s superpositions for analog implementation of the digit-wise computations in this larger base.

4 Conclusions

Arbitrary BFs can be implemented using:
- classical Boolean gates, but require exponential size;
- TGs, but (again) in exponential size (still, there are exponential gaps between classical Boolean solutions and TG ones);
- analog building blocks in linear size (having linear fan-in and polynomial precision weights and thresholds); the nonlinear activation function is the identity function.

The main conclusion is that size-optimal hardware implementations of BFs can be obtained only using analog (or mixed analog/digital) circuitry. The high precision required by the solution based on Kolmogorov’s superpositions can be tackled by decomposing a BF into simpler BFs. This is mathematically equivalent to computing in larger bases. Due to the reduced number of inputs, Kolmogorov’s superpositions can be used to design the analog implementations of the digit-wise computations in such larger bases.

References

Figure 3. The PARITY problem: (a) solution using COMPARATORS (inputs are translated by constants); (b) mixed analog/digital hand crafted solution (using COMPARATORS with constants and classical digital circuitry (there are direct connections between inputs and the nodes from the second layer).

[35] Y. Ito. Approximation of Functions on a Compact Set by Fi-


