LA-UR- 97-483

CONF-970598--1

Title: ENTROPY BASED COMPARISON OF NEURAL NETWORKS FOR CLASSIFICATION

Author(s): Sorin Draghici
Valeriu Beiu

APR 10 1997
OSTI

Submitted to: The 9-th Italian Workshop on Neural Nets
WIRN VIETRI-97
May 22-24, 1997
Vietri Sul Mare, Salerno Italy

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Los Alamos
NATIONAL LABORATORY

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. The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Form No. 836 R5
ST 2629 10/91

## 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, make 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.

# Entropy based comparison of neural networks for classification

**Sorin Draghici[1] and Valeriu Beiu[2*]**

[1]Vision and Neural Networks Laboratory,
Wayne State University,
431 State Hall, Detroit, 48202, MI, USA

[2]Los Alamos National Laboratory,
Division NIS-1, Mail Stop D466
Los Alamos, NM 87545, USA

**Abstract** - This paper discusses some entropy based bounds for the case of real and limited integer weights neural networks. It is shown that a neural network using real weights can solve a dichotomy of $m = m_+ + m_-$ patterns using a number of bits less than $\#bits < \max(m_+, m_-) \cdot n \left\{ \lceil \log(D/d) \rceil + 2 \right\}$ where n is the number of dimensions and D and d are the maximum and minimum distance between patterns in opposite classes, respectively. In the case of limited integer weights, it is shown that a neural network using integer weight in the range [-p,p] can solve a dichotomy of patterns in general positions with a number of bits greater than $\#bits > \min(m_+, m_-) \cdot n \cdot \lceil \log 2pD \rceil$.

## 1. Introduction

In recent years, multilayer feedforward neural networks (NN) have been shown to be very effective tools in many different applications [Arbib, 1995; Fiesler, 1996]. A natural and essential step in continuing the diffusion of these tools in our day by day use is their hardware implementation which is by far the most cost effective solution for large scale use. When the hardware implementation is contemplated, the issue of *the size* of the NN becomes crucial because the size is directly proportional with *the cost* of the implementation. In this light, any theoretical results which establish bounds on the size of a NN for a given problem is extremely important. In the same context, a particularly interesting case is that of the neural networks using limited integer weights. These networks are particularly suitable for hardware implementation because they need less space for storing the weights and the fixed point, limited precision arithmetic has much cheaper implementations in comparison with its floating point counterpart.

This paper presents an entropy based analysis which completes, unifies and correlates results partially presented in [Beiu, 1996, 1997a] and [Draghici, 1997]. Tight bounds for real and integer weight neural networks are calculated.

## 2. Previous results

The problem to find the smallest *size* NN which can realize an arbitrary function given a set of $m$ vectors (examples, or points) in $n$ dimensions is not new. Many results have been obtained for NNs having a *threshold* activation function. This is probably due to the fact that this line of research was continuing on the rigorous results already obtained in the literature dealing with threshold logic from the mid 60s (see references in [Beiu, 1997b; Fiesler, 1996]). Probably the first lower bound on the *size* of a threshold gate circuit for "almost all" n-ary Boolean functions was given by Neciporuk in 1964: $\text{size} \geq 2\left(2^n/n\right)^{1/2}$ [Neciporuk, 1964]. Later, Lupanov has proven a very tight upper bound for the case in which depth=4: $\text{size} \leq 2\left(2^n/n\right)^{1/2} * \left\{1 + \Omega\left[\left(2^n/n\right)^{1/2}\right]\right.$ [Lupanov, 1973]. Similar existence exponential bounds can be found in [Bruck, 1989], while in [Siu, 1991] a $\Omega\left(2^{n/3}\right)$ existence lower bound for arbitrary Boolean functions has been presented.

For classification problems, one of the first result was that a NN with only one hidden layer having m-1 nodes could compute an arbitrary dichotomy (sufficient condition). The main improvements are:

- Baum in [Baum, 1988] presented a NN with one hidden layer $\lceil m/n \rceil$ neurons capable of realizing an arbitrary dichotomy on a set of $m$ points *in general position* in $\mathbf{R}^n$; if the points are on the corners of the n-dimensional hypercube (i.e., binary vectors), m-1 nodes are still needed;

---

* On leave of absence from "Politehnica" University of Bucharest, Department of Computer Science, Spl. Independentei 313, RO-77206, Romania

- a slightly tighter bound was proven in [Huang, 1991]: only $\lceil 1+(m-2)/n \rceil$ neurons are needed in the hidden layer for realizing an arbitrary dichotomy on a set of $m$ points which satisfy a more relaxed topological assumption (only 'some' points are required to be *in general position*); also, the m-1 nodes condition was shown to be the least upper bound needed;
- Arai in [Arai, 1993] showed that m-1 hidden neurons are necessary for arbitrary separability (any mapping between input and output for the case of binary-valued units), but improved the bound for the two-category classification problem to *m/3* (without any condition on the inputs).

These results show that for binary inputs the *size* grows exponentially (as $m \le 2^n$). Some existence lower bounds for the arbitrary dichotomy problem are (see [Hassoun, 1995]):

- a *depth*-2 NN requires at least $m/\lfloor n\log(m/n) \rfloor$ hidden neurons (if $m \ge 3n$);
- a *depth*-3 NN requires at least $2(m/\log m)^{1/2}$ neurons in each of the two hidden layer (if $m >> n^2$); this bound is identical to the one presented in [Neciporuk, 1964] for $m = 2^n$;
- an arbitrarily interconnected NN without feedback needs $(2m/\log m)^{1/2}$ neurons (if $m >> n^2$).

Several other bounds for arbitrary Boolean functions can be found in [Paugam-Moisy, 1992]. All of these are: (i) revealing a gap between the upper and the lower bounds, thus encouraging research efforts to reduce (or even close) these gaps; (ii) suggesting that networks with more hidden layers might have a smaller *size*.

## 3. Theoretical considerations

### 3.1 Existence issues: Does a solution exist?

Before trying to establish any bounds for the necessary number of bits, one has to investigate whether the problem can be solved with the given tools. This issue is simpler in the case of the real weight networks. In this case, it is sufficient to show that a solution exists. Since the network uses real weights, if a solution exists, such a solution can be implemented[1] (because the hyperplanes can be placed in any position required by the solution). The issue becomes a little more complicated for the case of limited precision integer weights. In such situation, the hyperplanes are restricted to only certain discrete positions. The answers to these questions are given by the following propositions.

**Proposition 1**

A set of patterns from two classes in n dimensions for which the minimum distance between two patterns of opposite classes is *d*, can always be separated using a uniform quantization of the space with elementary hypercubes of side $d/(k\sqrt{n})$ where k>1.

The proof is constructive and can be found in [Beiu, 1996]. For space reasons, it will be omitted here. In [Beiu, 1996b], it is shown that the problem can always be solved even for the more convenient k=1.

**Proposition 2**

Using integer weights in the range [-p, p], one can correctly classify any set of patterns for which the minimum distance between two patterns of opposite classes is $d_{min}=1/p$.

The proof is based on induction on $p$ and on the number of dimensions $n$. Again, due to space limitations, the proof will be omitted here but can be found in [Draghici, 1997].

---

[1] Note that this paper does not deal with the *training* i.e. the existence of *an algorithm* able to find a given solution. We are only interested in the possibilities of *a network* (or architecture). *A network* can solve a problem if and only if a solution exists, independently of the existence of *an algorithm* able to find this solution.

### 3.2 Complexity issues: how big a network does one need in order to solve a given problem?

The interesting issue we have proposed to tackle is how complex should the network be for a given problem? Are there any bounds for the complexity of a network? In order to answer these questions, we must briefly discuss how we are going to measure the complexity of a network.

Many measures of complexity have been proposed. Among them, there are the *depth* (the number of edges on the longest input to the output path) and the *size* (the number of nodes). For VLSI implementation purposes, the *depth* can be put into correspondence with the delay and the *area* can be put into correspondence with the *area* of a VLSI chip. However, these measures are not the best criteria because the area of a neuron depends on the precision of its associated weights. Better criteria are the total *number of connections* [Hammerstrom, 1988; Abu-Mostafa, 1988; Klaggers, 1993; Phatak, 1994; Mason, 1995], the total *number-of-bits* needed to represent the weights [Bruck, 1990; Williamson, 1991] or *the sum of all the weights and thresholds* [Beiu, 1993, 1994, 1994b]. The total *number of bits* is discussed further in [Denker, 1988; Beiu, 1994] etc. Keeping in mind that our ultimate goal is to build hardware implementations, we will adopt the total *number of bits* as an appropriate measure for the complexity of a neural network.

We shall now consider the case of a set of patterns of two classes with the minimum distance between two patterns from opposite classes $d=d_{min}$ and we shall calculate some bounds on the number of bits necessary for both the case of real and limited integer weights.

**Proposition 3**

Let us consider the dichotomy of $m = m_+ + m_-$ examples from $\mathbf{R}^n$. Then, the number of bits necessary for the separation of the patterns (in general positions) using real weights is bounded by:

$$\#bits < \max(m_+, m_-) \cdot n\lceil \log(D/d) \rceil + 2 .$$

**Proof**

Find the examples (from the two classes) which are closest to one another: $x_d$, $y_d$ (the distance between them is d). Translate the origin in $x_d$ and rotate the axes such that $x_d$, $y_d$ become the opposite corners of a hypercube. The side length is $l = d/\sqrt{n}$ and we can use it as a step to quantize the whole space. As there are no patterns situated at a distance smaller than d, it is certain that no hypercube will contain examples from opposite classes.

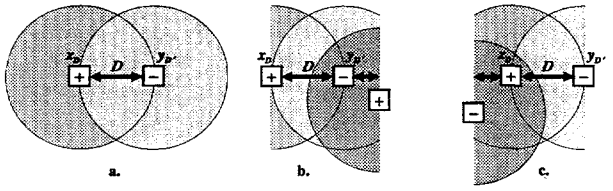


Fig. 1 Bounding the sub-space for a given set of examples

Because the largest distance between two patterns is D, there is a ball in $\mathbf{R}^n$ of radius D which contains all m examples (see Fig. 1). All the $m_+$ (respectively $m_-$) examples are contained inside a ball of radius D centered in that example from the opposite class which is used to determine the largest distance D: $y_D$ and $x_D$. Furthermore, there are only three possible cases:

- all the examples are in the sub-space determined by the intersection of two balls; all m examples are inside the intersection of two balls of radius D;
- there are positive examples situated on the other side (with respect to $x_D$) of the hyperplane orthogonal to the segment $x_Dy_D$ and containing $y_D$; select the furthest positive example and use it as the center of a third ball of radius D; all negative examples $m_-$ are inside a subspace of the intersection of two balls of radius D;
- there are negative examples situated on the other side (with respect to $y_D$) of the hyperplane orthogonal to the segment $x_Dy_D$ and containing $x_D$; now, all the $m_+$ examples are inside a sub-space of the intersection of two balls of radius D.

It is now clear that a bound on the number of bits can be obtained if one chooses to separate the least numerous class which contains minimum of $m_-$ and $m_+$ patterns.

From [Beiu, 1997], the volume of the intersection of two balls of radius r in $\mathbf{R}^n$, placed such that the center of each one is on the boundary of the other one, is $V(r,n) = 2\alpha(n-1)r^n \cdot a(n)$ where $\alpha(n) = \frac{\pi^{n/2}}{\Gamma(n/2+1)}$ is the volume of the unit ball in $\mathbf{R}^n$ and $a(n) = \frac{n-1}{n} \cdot a(n-2) - \frac{3^{(n-1)/2}}{n \cdot 2^n}$. Then, a bound on the number of bits can be calculated as $\#bits_{example} = \left\lceil \log\left( \frac{V(D,n)}{v_{hc}(d,n)} \right) \right\rceil = \left\lceil \log\left( \frac{2\pi^{(n-1)/2} D^n a(n) n^{n/2}}{d^n \Gamma[(n-1)/2+1]} \right) \right\rceil$.

From the same reference $\alpha(n) < \int_{\pi/6}^{\pi/2} \left(\sqrt{3}/2\right)^n d\theta = \left(\frac{\sqrt{3}}{2}\right)^n \cdot \frac{\pi}{3}$ and by using Stirling's formula $n! > \sqrt{2\pi n} \cdot (n/e)^n$ we obtain:

$$\#bits_{example} = \left\lceil \begin{array}{l} 1 + \frac{n}{2}\log\pi - \frac{1}{2}\log\pi + n\log(D/d) + \log\pi - \log 3 + \frac{n}{2}\log 3 - n + \frac{n}{2}\log n - \frac{1}{2}\log\pi - \\ - \frac{1}{2}\log(n-1) - \frac{n-1}{2}\log(n-1) + \frac{n-1}{2} + \frac{n-1}{2}\log e \end{array} \right\rceil$$

$$\#bits_{example} < \lceil n\log(D/d) + 1.8396n \rceil < n\{\lceil \log(D/d) \rceil + 2\}$$

We can choose M=max($m_-$, $m_+$) so we make sure that the result is independent of the case chosen in Fig. 1 and, by multiplication with M, the proof is concluded.

**Proposition 4**

Let us consider a set of $m = m_+ + m_-$ patterns of two classes in the hypersphere of radius D<=1 centred in origin of $\mathbf{R}^n$. Let us consider $d_{min}$=1/p the minimum distance between two patterns belonging to different classes. Then, the number of bits necessary for the separation of the patterns (in general positions) using weights in the set {-p, -p-1, ...,0, 1, ...,p} is bounded by

$$\#bits > \min(m_+, m_-) \cdot n \cdot \lceil \log 2pD \rceil$$

**Proof**

From proposition 2, it follows that one can divide the space using hyperplanes implemented with the given limited precision weights such that the maximum internal distance in any one region is less than 1/p.

The proof of proposition 2 is based on geometrical considerations which are illustrated in Fig. 2 and Fig. 3. Fig. 2 shows the hyperplanes which can be implemented with integer weights in the range [-3,3] (drawn in the square [-1,1]). There will be a certain number of 'large' volumes in which the maximum internal distance is 1/3-ε (1/p-ε in the general case) and a number of smaller volumes. Fig. 3 presents one of the 'large' volumes in 3D.

One can calculate the number of bits necessary for the representation of one example $p_i$ as

$$\#bits_{p_i} = \left\lceil \log \frac{V_{total}}{V_{iv}} \right\rceil$$

where $V_{total}$ is the total volume of the problem and $V_{iv}$ is the individual volume of the region which encloses (and separates) the pattern $p_i$. But all individual volumes are smaller or equal to the 'large' volumes which

have the maximum internal distance 1/p-ε (see [Draghici, 1997]). In turn, this volume is convex (from construction) and therefore smaller than the volume $V_{hs}$ of the hypersphere of diameter 1/p. Therefore:

$$\#bits_{p_i} = \left\lceil \log \frac{V_{total}}{V_{iv}} \right\rceil > \left\lceil \log \frac{V_{total}}{V_{hs}} \right\rceil$$

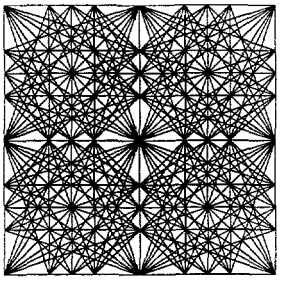

Fig. 2 The hyperplanes which can be implemented with integer weights in the range [-3,3] drawn in the square [-1,1].

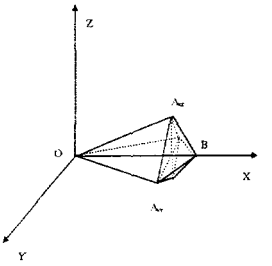


Fig. 3 The volume $V^3$ (in 3D). The largest distance between two points in this volume is OB along the x axis and is 1/p-ε.

For n even we have:

$$\#bits_{p_i} > \left\lceil \log \frac{V_{total}}{V_{hs}} \right\rceil = \left\lceil \log \frac{V_{total}}{\frac{\pi^{n/2}}{\Gamma(n/2+1)} d^n} \right\rceil = \left\lceil \log \frac{V_{total}}{\frac{\pi^{n/2}}{(n/2)!} d^n} \right\rceil$$

But, from hypothesis, all patterns are included in the sphere of radius D centered in the origin. Note that 1/2p<D<1. The first inequality comes from the fact that 1/p is the minimum distance in-between two patterns and 2D is the diameter of the hypersphere containing all patterns. The second one is necessary because proposition 1 was proved for the hypercube $[-1,1]^n$ . In these conditions, $V_{total}$ is the volume of the sphere of radius D and the bound can be written as:

$$\#bits_{p_i} > \left\lceil \log \frac{\frac{\pi^{n/2}}{(n/2)!} D^n}{\frac{\pi^{n/2}}{(n/2)!} d^n} \right\rceil = \left\lceil \log \frac{D^n}{d^n} \right\rceil = \left\lceil n \log \frac{D}{d} \right\rceil = = \left\lceil n \log \frac{D}{\frac{1}{2p}} \right\rceil = \lceil n \log 2pD \rceil$$

By multiplying with $\min(m_+, m_-)$ i.e. the number of patterns in the smallest class, we obtain $\#bits > \min(m_+, m_-) \cdot n \cdot \lceil \log 2pD \rceil$. A similar expression can be obtained for n odd. QED.

This lower bound must not be interpreted as an absolute one. For a particular problem, when the patterns are in particularly favourable positions, more than one pattern from the same class can share the same volume and thus, the number of bits can be further reduced.

The upper bound holds even for the limited integer weight case. Therefore, for this case:

$$\min(m_+, m_-) \cdot n \cdot \left\{ \left\lceil \log\left(\frac{D}{d}\right) \right\rceil \right\} < \#bits < \max(m_+, m_-) \cdot n \cdot \left\{ \left\lceil \log\left(\frac{D}{d}\right) \right\rceil + \frac{5}{2} \right\}.$$

## 4. Conclusions

This paper has addressed two issues. Firstly, it has been shown that both in the case of real and limited integer weights, the problem can always be solved. Then, based on the entropy of the data set, tight lower and upper bounds for the case of real and limited integer weights neural networks have been calculated.

Although the proof for the given bounds is constructive, the shape of the bounding space can not be easily used in practice because it is too expensive computationally. In practice, the simplest bounding space is a hypercube. By taking a hypercube of side lenght 2D, the problem can be solved but the upper bound becomes $m\left\lceil \log \frac{\left(2D^n\right)}{\left(d/\sqrt{n}\right)^n} \right\rceil < m\lceil 5n/2 + n\log(D/d) + (n\log n)/2 \rceil = O(mn\log n)$. In other words, this has lead to a logarithmic increase on the upper bound.

## 5. Bibliography


[Abu-Mostafa,1988]-Abu-Mostafa Y.S., Connectivity vs. Entropy, NIPS'87, D.Z. Anderson(Ed), Amer. Inst. Of Phys., NY,1988,1-8

[Arai, 1993]-Arai, M. Bounds on the number of hidden units in binary-valued three-layer neural networks, NN **6** (6) (1993) 855-860.

[Arbib, 1995] - Arbib, M.A. (ed.): The handbook of brain theory and neural networks, MIT Press, Cambridge, MA (1995).

[Baum, 1988] - Baum, E.B.: On the Capabilities of Multilayer Perceptrons, J. of Complexity **4** (1988) 193-215.

[Beiu, 1993] - Beiu, V., Peperstraete, J., Vandewalle, J., Lauwereins, R., Comparison and threshold gate decomposition, in Myers, D.J., Murray, A.F.: MicroNeuro'93 (Edinburgh, UK), UnivEd Tech. Ltd., Edinburgh, 83-90, 1993

[Beiu, 1994] - Beiu V. - Neural Networks Using Threshold Gates: A Complexity Analysis of Their Area and Time Efficient VLSI Representations, Ph.D. thesis, Katholieke Universiteit Leuven, 1994.

[Beiu, 1994b] - Beiu, V., Peperstraete, J., Vandewalle, J., Lauwereins, R., Area-time performnaces of some neural computations, in Borne, P., Fukuda, T., Tzafestas, S.G., Symp. On Signal Processing, Robotics and Neural Networks, GERF EC, Lille, 664-668, 1994

[Beiu, 1996] - Beiu, V., Entropy bounds for classification algorithms, Neural Network World, **6**, No. 4, 497-505, IDG Press, 1996

[Beiu, 1996b] - Beiu V. ,Taylor J.G., Direct synthesis of neural networks, Proc. MicroNeuro'96 (Lausanne, Switzerland), IEEE CS Press, Los Alamitos, CA, 257-264, 1996

[Beiu, 1997a] - Beiu, V., T. de Pauw, Tight bounds on the size of neural networks for classification problems, subm. for IWANN'97

[Beiu, 1997b] - Beiu, V.,VLSI complexity of discrete neural networks, Gordon and Breach (1997).

[Bruck, 1989] - Bruck, J., Smolensky, R.: Polynomial threshold functions, $AC^0$ functions and spectral norms. Res. Rep. RJ 7410 (67387), 11/15/89, IBM Yorktown Heights, NY (1989). Also in SIAM J. Computing 21(1) (1992) 33-42.

[Bruck, 1990] - Bruck J., Goodman J.W., On the power of neural networks for solving hard problems, NIPS'87, D.Z. Anderson (Ed.), Amer. Inst. Of Phys., NY, 1988, 137-143 (also in J. of Complexity, **6**, 1990, 129-135)

[Denker, 1988] - Denker J.S., Wittner B.S., Network Generality, Training Required and Precision Required, NIPS'88, D.Z. Anderson (Ed.), Amer. Inst of Phys., New York, 1988, 219-222

[Draghici, 1997] - Draghici S., Sethi, I.K: On the possibilities of the limited precision weights neural networks in classification problems, submitted for International Work-Conference on Artificial and Natural Neural Networks IWANN'97

[Fiesler, 1996] - Fiesler, E., Beale, R. (eds.): Handbook of neural computation, Oxford University Press and the Institute of Physics Publishing, NY (1996), Parts E1.4: Digital Integrated Circuits and G: Neural Networks in Practice: Case Studies

[Hammerstrom, 1988] - Hammerstrom D., The connectivity analysis of simple associations - or - How many connections do you need?, NIPS'87, D.Z. Anderson (Ed.), Amer. Inst. Of Phys., New York, 1988, 338-347

[Hassoun, 1995] - Hassoun, M.H.: Fundamentals of artificial neural networks, MIT Press, Cambridge, MA (1995).

[Huang, 1991] - Huang, S.-C., Huang, Y.-F.: Bounds on the number of hidden neurons of multilayer perceptrons in classification and recognition, IEEE Trans. on Neural Networks **2** (1) (1991) 47-55.

[Klaggers, 1993] - Klaggers H., Soegtrop M., Limited fan-in random wired cascade-correlation learning architecture, MicroNeuro'93, D.J. Myers and A.F.Murray (Eds.), Univ.Ed Tech. Ltd. Edinburgh, 1993, pp. 79-82

[Lupanov, 1973] - Lupanov, O.B.: The synthesis of circuits from threshold elements, Problemy Kibernetiki 20 (1973) 109-140.

[Mason, 1995] - Mason R.D., Robertson W., Mapping Hierarchical Neural Networks to VLSI Hardware, Neural Networks, **8**, 6, 1995, 905-913

[Neciporuk, 1964b] - Neciporuk, E.I.: The synthesis of networks from threshold elements, Problemy Kiber-netiki **11** (1964) 49-62. English translation in Automation Express **7** (1) 35-39 and **7** (2) 27-32.

[Paugam-Moisy, 1992] - Paugam-Moisy, H.: Optimisation des réseaux des neurones artificiels. Ph.D. thesis, Laboratoire de l'Informatique du Parallélisme LIP-IMAG, École Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon, France (1992).

[Phatak, 1994] - Phatak D.S., Koren I., Connectivity and performance tradeoffs in the cascade-correlation learning architecture, IEEE Trans. NN's, 5, 6, 1994, 930-935

[Siu, 1991] - Siu, K.-Y., Roychowdhury, V., Kailath, T.: Depth-size tradeoffs for neural computations, IEEE Trans. on Comp., **40** (12) (1991) 1402-1412.

[Williamson, 1991] - Williamson R.C., Entropy and the complexity of feedforward neural networks, NIPS'90, R.P. Lippmann, J.E.Moody and D.S. Touretzky (Eds.), Morgan Kaufmann, San Mateo, 1991, pp. 946-952