RUTCOR
RESEARCH
REPORT

Bounds on the Degree of High Order Binary Perceptrons

Eddy Mayoraz[a]

RRR 44-95, December 1995

RUTCOR • Rutgers Center for Operations Research • Rutgers University • P.O. Box 5062 • New Brunswick New Jersey • 08903-5062
Telephone: 908-445-3804
Telefax: 908-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/ rrr

[a]IDIAP Institut Dalle Molle d'Intelligence Artificielle Perceptive c.p. 592, CH-1920 Martigny, Switzerland. email: mayoraz@idiap.ch

Rutcor Research Report

RRR 44-95, December 1995

Bounds on the Degree of High Order Binary Perceptrons

Eddy Mayoraz

Abstract. High order perceptrons are often used in order to reduce the size of neural networks. The complexity of the architecture of a usual multilayer network is then turned into the complexity of the functions performed by each high order unit and in particular by the degree of their polynomials. The main result of this paper provides a bound on the degree of the polynomial of a high order perceptron, when the binary training data result from the encoding of an arrangement of hyperplanes in the Euclidian space. Such a situation occurs naturally in the case of a feedforward network with a single hidden layer of first order perceptrons and an output layer of high order perceptrons. In this case, the result says that the degree of the high order perceptrons can be bounded by the minimum between the number of inputs and the number of hidden units.

Acknowledgements: This work was initiated while I was a postdoctoral fellow, supported by RUTCOR and DIMACS—Center for Discrete Mathematics and Theoretical Computer Science. I am particularly thankful to my colleague, Dr. Motakuri Ramana, for helpful discussions on this matter.

1 Introduction

The usage of a single perceptron is very restricted, since limited to tasks which are linearly separable. To extend the computational power of this model we can either introduce hidden layers of non-linear functions or increase the possibilities of the perceptron by replacing its linear combination of the inputs by a polynomial combination. The latter option leads to a new model of unit called *high order perceptron*.

The computational power of a polynomial is such that neural networks with high order perceptrons can be resumed to networks of a single unit. However, to restrict the computational complexity as well as for the sake of the generalization, it is essential to limit the complexity of the polynomial. This can be done by bounding either the degree of the polynomial or its number of terms.

In the case of binary inputs coded as -1 and $+1$, there is no use to take a power of a variable, since $x^k = \pm x \; \forall k > 0$ when $x \in \{-1, +1\}$, and thus the degree of any polynomial over n variables is at most n. Moreover, for any Boolean function $f : \{-1, +1\}^n \rightarrow \{-1, +1\}$, the *spectral representation*, well known in harmonic analysis [3], is the unique polynomial of the form:

$$f(\boldsymbol{b}) = \sum_{K \subseteq \{1,\ldots,n\}} w_K \prod_{k \in K} b_k, \tag{1}$$

where the sum is taken over the 2^n possible subsets K of $\{1, \ldots, n\}$. The number of non-zero terms of this polynomial being usually exponential, it has to be limited to make this model suitable for applications.

The computational power of feedforward networks where each unit is a high order perceptron with a number of terms polynomial bounded in n, the number of inputs, has been investigated in [1]. In this work we show how some *a priori* knowledge on a given set of binary data can be used to bound the degree of the polynomial of a single high order perceptron, while guaranteeing a correct learning of the data.

Binary data appears rarely as such in the real world. Quite often they results from a preprocessing of data of a more general nature (*e.g.* continuous or nominal) expressed through logical predicates (*e.g.* "is older than 50", "is red", "has a rent higher than 0.27 times his income"). The main result of this paper states that in such circumstances, the degree of the polynomial in (1) can be bounded by the size of the space of the original data.

A feedforward network with d inputs of continuous activations, one hidden layer of n usual perceptrons a high order perceptron as output, presents a particular situation of this type. The input space is continuous, let say $\mathbb{R}^d$, and the hidden layer provides a mapping of $\mathbb{R}^d$ into $\{-1, +1\}^n$. Our main theorem will imply that the degree of the output unit can always be bounded by $\min\{d, n\}$.

2 Definitions and Notations

A *Boolean function* is a mapping $f : \{-1, +1\}^n \rightarrow \{-1, +1\}$. A *partial Boolean function* is a mapping $f : D \subseteq \{-1, +1\}^n \rightarrow \{-1, +1\}$, and D is called the *domain of f*. An *extension* of a partial Boolean function f is any Boolean function that coincides with f on its domain. The *spectral representation* of a Boolean function f is the polynomial of the form (1).

For any subset X of the Euclidian space $I\!R^d$, X^c denotes the complement $I\!R^d \backslash X$, while $\overline{X}$ and X^o denote respectively the closure of X and the border of X (*n.b.* $X^o = \overline{X} \cap \overline{X^c}$) , according to the usual topology of the Euclidian space. In $I\!R^d$, a *closed half-space* of paremeters $\boldsymbol{w} \in I\!R^d$ and $w_0 \in I\!R$ is the set $\{\boldsymbol{x} \in I\!R^d \mid \boldsymbol{x}^\top \boldsymbol{w} \geq w_0\}$. If H is a closed half-space of $I\!R^d$ of parameters $\boldsymbol{w}$ and w_0, H^o clearly denotes the hyperplane $\{\boldsymbol{x} \in I\!R^d \mid \boldsymbol{x}^\top \boldsymbol{w} = w_0\}$. A finite subset of closed half-spaces in $I\!R^d$ is called an *arrangement* of $I\!R^d$. A *cell* of an arrangement $\mathcal{H} = \{H_1, \ldots, H_n\}$ of $I\!R^d$ is a non-empty intersection of n half-spaces among $H_1, H_1^c, \ldots, H_n, H_n^c$. The set of cells of an arrangement $\mathcal{H}$ is denoted $\mathcal{C}_\mathcal{H}$.

An arrangement $\mathcal{H} = \{H_1, \ldots, H_n\}$ of $I\!R^d$ is in *general position* if

$$\begin{aligned} & \cap_{k\in K} H_k^o = \emptyset \quad \forall K \subseteq \{1, \ldots, n\}, |K| > d \; ; && (2) \\ \text{and} \quad & \cap_{k\in K} H_k^o \neq \emptyset \quad \forall K \subseteq \{1, \ldots, n\}, |K| \leq d \; . && (3) \end{aligned}$$

Condition (2) states that there are never more than d hyperplanes through the same point, while condition (3) implies, among others, that there are no parallel hyperplanes. The number of cells of an arrangement of n half-spaces in general position in $I\!R^d$ is given by the well-known formula

$$N_n^d = \sum_{i=0}^{\min\{d,n\}} \binom{n}{i} \; , \tag{4}$$

which can be easily proved by induction and which is attributed to Ludwig Schläfli, a Swiss mathematician of the 19^{th} century.

To any arrangement $\mathcal{H} = \{H_1, \ldots, H_n\}$, we can associate an injective mapping $\phi_\mathcal{H} : \mathcal{C}_\mathcal{H} \to \{-1, +1\}^n$: the Boolean vector $\boldsymbol{b} = \phi_\mathcal{H}(C)$ is such that $b_k = 1$ if $C \subseteq H_k$ and $b_k = -1$ if $C \subseteq H_k^c$. Let us denote $D_\mathcal{H}$ the subset $\phi_\mathcal{H}(\mathcal{C}_\mathcal{H})$ of $\{-1, +1\}^n$. To each bipartition of the cells of an arrangement $\mathcal{H}$ corresponds a partial Boolean function defined from $D_\mathcal{H} \subseteq \{-1, +1\}^n$ to $\{-1, +1\}$.

3 Main results

The main result presented in theorem 3.1 has been proposed independently by L. Gurvits [2] in a slightly more general setting based on the VC-Dimension of classes of discriminators. However, the algebraic approach used here to prove this result leads us to a stronger statement expressed in theorem 3.2, which is completely new, to the best of our knowledge.

> **Theorem 3.1** For an arrangement $\mathcal{H}$ of n half-spaces in $I\!R^d$, any partial Boolean function $f : D_\mathcal{H} \subseteq \{-1, +1\}^n \to \{-1, +1\}$ can be exactly represented by a polynomial of degree $\leq \min\{d, n\}$.

Proof: Since the spectral representation (1) of a Boolean function is of degree at most n, we only have to show that a function defined on $D_\mathcal{H}$ has an extension whose spectral representation is bounded by d.

Let us first prove this result when the additional assumption of general position is made on the arrangement $\mathcal{H}$. An easy argument will then imply the result for arbitrary arrangements.

Let $\mathcal{P}_n^d$ denote the set of all subsets of $\{1, \ldots, n\}$ of cardinality at most d. Note that $|\mathcal{P}_n^d| = N_n^d$. Let $\boldsymbol{A}_n^d$ be the $N_n^d \times N_n^d$-matrix, with rows indexed by the Boolean vectors of $D_{\mathcal{H}}$, with columns indexed by the elements of $\mathcal{P}_n^d$ and with ± 1 coefficients defined as follows:

$$a_{b,K} \quad = \quad \prod_{k \in K} b_k, \quad \forall \boldsymbol{b} \in D_{\mathcal{H}}, \quad \forall K \in \mathcal{P}_n^d \ .$$

To a partial Boolean function $f : D_{\mathcal{H}} \to \{-1, +1\}$, let us associate the Boolean vector $\boldsymbol{f} \in \{-1, +1\}^{D_{\mathcal{H}}}$ indexed by the elements of $D_{\mathcal{H}}$ and such that $f_b = f(\boldsymbol{b})$. Similarly, let us specify a spectral representation of the form (1) and of degree d by a vector $\boldsymbol{w} \in I\!R^{\mathcal{P}_n^d}$ indexed by $\mathcal{P}_n^d$. A necessary and sufficient condition for a partial Boolean function $f : D_{\mathcal{H}} \to \{-1, +1\}$ to have an extension with a spectral representation of degree at most d is that the system

$$\boldsymbol{A}_n^d \boldsymbol{w} = \boldsymbol{f} \tag{5}$$

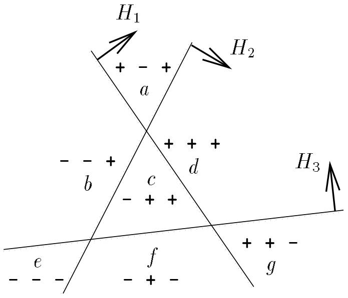
has a solution in $I\!R^{\mathcal{P}_n^d}$.

This system has a solution if matrix $\boldsymbol{A}$ is non-singular. This will be established by induction on the number n of half-spaces in $\mathcal{H}$.

$\boldsymbol{A}_0^d$ as well as $\boldsymbol{A}_n^0$ are equal to the 1×1 matrix $(+1)$ which is non-singular.

To establish an inductive relation between $\boldsymbol{A}_n^d$ and $\boldsymbol{A}_{n-1}^d$, consider the partition of $\mathcal{C}_{\mathcal{H}}$ into $\mathcal{C}_1 \uplus \mathcal{C}_2 \uplus \mathcal{C}_3$:

$$\begin{aligned}
\mathcal{C}_1 &= \{C \in \mathcal{C}_{\mathcal{H}} \mid \overline{C} \cap H_n^o = \emptyset\}, \\
\mathcal{C}_2 &= \{C \in \mathcal{C}_{\mathcal{H}} \mid C \cap H_n^o \neq \emptyset\}, \\
\mathcal{C}_3 &= \{C \in \mathcal{C}_{\mathcal{H}} \mid \overline{C} \cap H_n^o \neq \emptyset \ \text{ and } \ C \subseteq H_n^c\}.
\end{aligned}$$



$$\boldsymbol{A}_3^2 = \begin{array}{c} \\ \end{array}
\begin{array}{ccccccc|l}
\emptyset & \{1\} & \{2\} & \{1,2\} & \{3\} & \{1,3\} & \{2,3\} & \\
+ & + & - & - & + & + & - & a \ \mathcal{C}_1 \\
+ & - & - & + & + & - & - & b \\
+ & - & + & - & + & - & + & c \ \mathcal{C}_2 \\
+ & + & + & + & + & + & + & d \\
+ & - & - & + & - & + & + & e \\
+ & - & + & - & - & + & - & f \ \mathcal{C}_3 \\
+ & + & + & + & - & - & - & g
\end{array} \quad \mathcal{C}_{\mathcal{H}}$$

Figure 1: Simple illustration of the notations and the construction used.

Each cell C of the arrangement $\mathcal{H} = \{H_1, H_2, H_3\}$ in $I\!R^2$ is denoted by a letter from a to g and by $\phi_{\mathcal{H}}(C)$, where $+$ and $-$ stand for $+1$ and -1. The corresponding matrix $\boldsymbol{A}_3^2$ is presented on the right with its block structure.

This partition of $\mathcal{C}_{\mathcal{H}}$ induces a partition of $D_{\mathcal{H}}$ as $D_1 \uplus D_2 \uplus D_3$, where $D_i = \phi_{\mathcal{H}}(\mathcal{C}_i), i = 1, 2, 3$. If $\mu_k : \{-1, +1\}^n \to \{-1, +1\}^n$ denotes the application that inverses the k^{th} bit of a Boolean vector, D_i can be caracterized as follows:

$$\begin{aligned}
D_1 &= \{\boldsymbol{b} \in D_{\mathcal{H}} \mid \mu_n(\boldsymbol{b}) \notin D_{\mathcal{H}}\},\\
D_2 &= \{\boldsymbol{b} \in D_{\mathcal{H}} \mid \mu_n(\boldsymbol{b}) \in D_{\mathcal{H}} \text{ and } b_n = +1\},\\
D_3 &= \{\boldsymbol{b} \in D_{\mathcal{H}} \mid \mu_n(\boldsymbol{b}) \in D_{\mathcal{H}} \text{ and } b_n = -1\}.
\end{aligned}$$

By reordering its rows and columns, $\boldsymbol{A}_n^d$ can be expressed by the block structure

$$\begin{pmatrix} B & E \\ C & F \\ D & G \end{pmatrix},$$

where the sets of rows of B and E, C and F and D and G are indexed by D_1, D_2 and D_3 respectively; and the sets of columns of B, C and D and E, F and G are indexed by $\mathcal{P}_{n-1}^d$ and $\mathcal{P}_n^d \backslash \mathcal{P}_{n-1}^d$ respectively.

Observe that on the one hand, there is a one-to-one mapping between $\mathcal{C}_1 \cup \mathcal{C}_2$ and $\mathcal{C}_{\mathcal{H}'}$, where $\mathcal{H}' = \{H_1, \ldots, H_{n-1}\}$. Consequently,

$$\begin{pmatrix} B \\ C \end{pmatrix} = \boldsymbol{A}_{n-1}^d.$$

On the other hand, any cell in $\mathcal{C}_2$ has one face in H_n and thus there is also a one-to-one mapping between $\mathcal{C}_2$ and the set of cells of the arrangement $\{H_1 \cap H_n, \ldots, H_{n-1} \cap H_n\}$ in the $(d-1)$-dimensional subspace H_n. After removing n from each element of $\mathcal{P}_n^d \backslash \mathcal{P}_{n-1}^d$ we get $\mathcal{P}_{n-1}^{d-1}$, and thus $F = \boldsymbol{A}_{n-1}^{d-1}$.

By definition of D_2 and D_3, $D_3 = \mu_n(D_2)$ and thus, by reordering rows in D and G if necessary, we have

- $C = D$, since $n \notin K$ $\forall K$ indexing the columns of C and D;
- $F = -G$, since $n \in K$ $\forall K$ indexing the columns of F and G.

The determinant of a matrix does not change by reordering rows and columns or by replacing a row by a linear combination of this row with others. Thus

$$\begin{aligned}
\det(\boldsymbol{A}_n^d) &= \det \begin{pmatrix} B & E \\ C & F \\ C & -F \end{pmatrix} = \det \begin{pmatrix} B & E \\ C & F \\ 0 & 2F \end{pmatrix}\\
&= \det \begin{pmatrix} B \\ C \end{pmatrix} \det(2F) = \det(\boldsymbol{A}_{n-1}^d) \det(2\boldsymbol{A}_{n-1}^{d-1}) .
\end{aligned}$$

This recursive relation together with the initial statement on the non-singularity of $\boldsymbol{A}_n^0$ and $\boldsymbol{A}_0^d$ implies the non-singularity of $\boldsymbol{A}_n^d$ for any $d, n \geq 0$.

To generalize the proof to arbitrary arrangements (not necessarily in general position), it suffices to observe that an arrangement $\mathcal{H}$ can always be slightly modified to provide an arrangement $\mathcal{H}'$ in general position and such that there is a one-to-one mapping between $\mathcal{C}_{\mathcal{H}}$ and a subset of $\mathcal{C}_{\mathcal{H}'}$. Thus a system of the form (5) based on $\mathcal{H}$ will be a subsystem of the one based on $\mathcal{H}'$ and if the latter has a solution, so will the former. $\triangle$

Theorem 3.2 For an arrangement $\mathcal{H}$ of n half-spaces in $I\!R^d$, any partial Boolean function $f : D_{\mathcal{H}} \subseteq \{-1,+1\}^n \rightarrow \{-1,+1\}$ can be exactly represented by a polynomial containing a term $\prod_{k\in K} b_k$ only for the $K \subseteq \{1,\ldots,n\}$ such that $|K| \leq d$ and $\cap_{k\in K} H_k^o \neq \emptyset$.

Proof: Following the same line than in the previous proof, we can show that $\det(\boldsymbol{A}_{\mathcal{H}}) = \det(\boldsymbol{A}_{\mathcal{H}'})\det(\boldsymbol{A}_{\mathcal{H}''})$, where $\boldsymbol{A}_{\{X_1,\ldots,X_s\}}$ is a matrix defined in a similar way than A_n^d, with one row per cell in the arrangement $\{X_1,\ldots,X_s\}$ and one column per non empty intersection $\cap_{k\in K} X_k^o$; and where $\mathcal{H}$ denotes an arbitrary arrangement $\{H_1,\ldots,H_n\}$ in $I\!R^d$; $\mathcal{H}'$ is the arrangement $\{H_1,\ldots,H_{n-1}\}$ in $I\!R^d$ and $\mathcal{H}''$ is the arrangement $\{H_1\cap H_n,\ldots,H_{n-1}\cap H_n\}$ in the $(d-1)$-dimensional space H_n. $\triangle$

4 Simple usage of the results

The two theorems presented in this paper provide ways to bound the complexity of the polynomial of a high order perceptron, when the data that as to be learned is Boolean and results from an arrangement of hyperplanes in a low dimensional Euclidian space.

To illustrate the interest of these theorems, let us consider a board of go which is a grid of 19 rows and 19 columns. A compact binary encoding of each position requires at least $\lceil\log(19^2)\rceil = 9$ bits. Any arbitrary subset of the $19^2 = 361$ positions can be modeled by a Boolean function and harmonic analysis tells us that this function has a spectral representation of the form (1) with $2^9 = 512$ terms. On the contrary, a more natural encoding of each position on 36 bits is provided by the introduction of 36 separating lines in the plane (18 horizontal and 18 vertical). Any arbitrary subset of the positions of the board is now modeled as a partial Boolean function of 36 arguments, and the spectral representation would lead to a polynomial with 2^{36}, which is a lot if we ignore that many of them will have a zero coefficient for any subset of positions.

Theorems 3.1 and 3.2 precisely inform us about terms of the spectral representation that will always have a zero coefficient. The 361 Boolean vectors come from an arrangement embedded in $I\!R^2$ and by theorem 3.1 there exists such a polynomial of degree 2, *i.e.* with at most $N_{36}^2 = 1+36+630 = 667$ terms. Because most of these lines are parallel, by theorem 3.2 we know that there is a polynomial of degree 2 with at most $1+36+18^2 = 361$ terms. Thus, the second encoding is more practical since it has an easy geometrical interpretation, and it leads to a smaller polynomial.

It is interesting to note that this gain in the number of terms, from 512 to 361, is only due to the fact that 19^2 is not a power of 2. For a board with $n = 2^k$ positions (*e.g.* a chess board), the "geometrical" encoding of the board with the help of theorem 3.2, and the "compact" encoding would lead to two polynomials based on completely different sets of variables, but with the same number n of terms, non-zero in the spectral representation of any subset of the positions. Moreover, this analogy can be generalized to rectangular grids of arbitrary dimensions.

References

[1] Jehoshua Bruck and Roman Smolensky. Polynomial threshold functions, AC^0 functions and spectral norms. *SIAM J. Comput.*, 21(1):33–42, 1992.

[2] Leonid Gurvits. Some combinatorial and topological properties of perceptron networks. Oral communication at a DIMACS Workshop on Mathematical Theory of Neural Networks, July 1994.

[3] R. J. Lechner. Harmonic analysis of switching functions. In A. Mukhopadhyay, editor, *Recent Development in Switching Theory*. Academic Press, New York, 1971.