LETTER **Communicated by Michael Schmitt**

An Upper Bound on the Minimum Number of Monomials Required to Separate Dichotomies of $\{-1, 1\}^n$

Erhan Oztop
erhan@atr.jp
JST-ICORP Computational Brain Project and ATR Computational Neuroscience Laboratory, 2-2-2 Hikari-dai Soraku-gun, Kyoto, 619-0288, Japan

It is known that any dichotomy of $\{-1, 1\}^n$ can be learned (separated) with a higher-order neuron (polynomial function) with 2^n inputs (monomials). In general, less than 2^n monomials are sufficient to solve a given dichotomy. In spite of the efforts to develop algorithms for finding solutions with fewer monomials, there have been relatively fewer studies investigating maximum density ($\Pi(n)$), the minimum number of monomials that would suffice to separate an arbitrary dichotomy of $\{-1, 1\}^n$. This article derives a theoretical (upper) bound for this quantity, superseding previously known bounds. The main theorem here states that for any binary classification problem in $\{-1, 1\}^n$ ($n > 1$), one can always find a polynomial function solution with $2^n - 2^n/4$ or fewer monomials. In particular, any dichotomy of $\{-1, 1\}^n$ can be learned by a higher-order neuron with a fan-in of $2^n - 2^n/4$ or less. With this result, for the first time, a deterministic ratio bound independent of n is established as $\Pi(n)/2^n \leq 0.75$. The main theorem is constructive, so it provides a deterministic algorithm for achieving the theoretical result. The study presented provides the basic mathematical tools and forms the basis for further analyses that may have implications for neural computation mechanisms employed in the cerebral cortex.

1 Introduction

Higher-order neurons (units) or sigma pi units are computationally powerful extensions of linear neuron models (Rumelhart, Hinton, & Williams, 1986; Giles & Maxwell, 1987; Schmitt, 2005). These units capture the nonlinearity in the input-output relation through products of input variables called monomials. The net input to a higher-order unit is the sum of the monomials weighted by adjustable parameters. The output is obtained by the application of a predefined activation function (e.g., sigmoidal function or a threshold function) to the net input. When the output is a threshold function, these units are sometimes called polynomial threshold units. Accumulating biological data suggest that specific neurons in the cerebral

cortex compute in a multiplicative way (see Schmitt, 2002). Therefore, higher-order units whose monomials capture the nonlinear dendritic information processing can be considered better models for real neurons compared to the McCulloch-Pitts model (Mel & Koch, 1990; Mel, 1994). It is well known that the use of higher-order units increases the computational power and storage capacities of neural networks (Schmitt, 2002); however, the combinatorial growth in the number of monomials required for a given problem limits their application. Some work has been devoted to developing algorithms for finding a reduced set of monomials to realize a given classification problem without suffering from the combinatorial growth problem (e.g., Ghosh & Shin, 1992; Guler, 2001). Theoretical results concerning the bounds for Vapnik-Chervonenkis dimension of higher-order neurons have also been obtained (Schmitt, 2002, 2005). More relevant to this article is the study of the so-called polynomial threshold density of Boolean functions (i.e., dichotomies) (see, e.g., Saks, 1993), which indicates the minimum number of monomials over all the polynomial functions that solve a given classification problem in $\{-1, 1\}^n$. It is of both practical and theoretical importance to determine the maximum density, $\Pi(n)$, the number of monomials that one can always separate any dichotomy of $\{-1, 1\}^n$. Spectral theory of Boolean functions produced important results and elegant methods to derive some bounds on the maximum density, $\Pi(n)$. The best-known lower bound is 0.11×2^n (see, Saks, 1993; O'Donnell & Servedio, 2003). The upper bound is due to Gotsman (1989), who proved that the maximum density is at most $2^n - \sqrt{2^n} + 1$. These bounds tell us that every dichotomy of $\{-1, 1\}^n$ can be separated with $2^n - \sqrt{2^n} + 1$ or fewer monomials, and there are dichotomies that cannot be separated with fewer than 0.11×2^n monomials. In fact, it is known that there exist dichotomies that cannot be separated with a polynomially (in n) bounded number of monomials (Bruck, 1990). Recently, O'Donnell and Servedio (2003) improved the upper bound asymptotically to $2^n - 2^n/\mathrm{O}(n)$. This article further improves the latter bound by proving $\Pi(n) \leq 2^n - 2^n/4$, thereby, for the first time, establishing a deterministic ratio bound independent of n: $\Pi(n)/2^n \leq 0.75$.

1.1 A Motivating Example. Consider the dichotomy (fully specified binary classification problem) given in Table 1. For simplicity we use -1 and $+1$ for the class labels. A solution to this problem would be a polynomial of x_0, x_1, and x_2 with no powers greater than 1 (higher powers are not needed since $x_k^2 = 1$) such that the sign of the polynomial function evaluated at each x_0, x_1, x_2 picked from the rows of Table 1 coincides with the class label given in that row. There are infinitely many such polynomials. Here are some examples:

Table 1: Class Assignment Table for the Example of Section 1.1.

x_0	x_1	x_2	Class	p_1	p_2	p_3
1	1	1	−1	−4	−3	−1
−1	1	1	1	4	3	3
1	−1	1	1	4	1	3
−1	−1	1	−1	−4	−1	−5
1	1	−1	−1	−4	−1	−3
−1	1	−1	1	4	5	1
1	−1	−1	1	4	3	1
−1	−1	−1	1	4	1	1

Notes: The last three columns enlist the outputs of the polynomials. The sign of the values in these columns match the class labels, which indicates that all three polynomials are solutions to the given problem.

$$p_1 = -x_2x_1x_0 + x_2x_1 + x_2x_0 - x_2 - 3x_1x_0 - x_1 - x_0 + 1$$
$$p_2 = -x_2 - 2x_1x_0 - x_0 + 1$$
$$p_3 = -x_2x_1x_0 + x_2x_1 + x_2x_0 - 2x_1x_0.$$

It can be verified that these are solutions to the dichotomy (see the last three columns of Table 1). Note that the polynomial p_1 contains eight monomials, whereas p_2 and p_3 contains four monomials. One wonders whether it is possible to find a solution with fewer terms (monomials). The study presented in this article is motivated by this question. More generally, we pursue an answer to the question, "Can we find a general upper bound on the minimum number of monomials that one can separate any dichotomy of $\{-1, 1\}^n$?" The next section presents definitions needed for the derivations leading to an (affirmative) answer to this question.

2 Definitions

Definition 1. *A binary classification problem $C_n = (S_n^+, S_n^-)$ in $\{-1, 1\}^n$ is defined with two disjoint sets of input vectors $S_n^+ \subset \{-1, 1\}^n$ and $S_n^- \subset \{-1, 1\}^n$. We use Λ_n to represent the collection of all dichotomies (i.e., fully specified binary classification problems) in $\{-1, 1\}^n$. Note that $|\Lambda_n| = 2^{2^n}$. When it is clear from the context, the subscript n may be suppressed.*

Definition 2. *A polynomial function (of dimension n) is a polynomial over the field of real numbers interpreted as a function of $\{-1, 1\}^n$. We represent the set of polynomial functions of dimension n with $\Theta_n = \{p(x) \in \Re[x] | p(x) : \{-1, 1\}^n \to \Re\}$.*

Definition 3. *Any polynomial function, $p(x_0, x_1, \ldots, x_{n-1})$ can be written as $\sum_{i=1}^{2^n} a_i \prod_{k \in S_i} x_k$ (i runs through all the subsets, $S_i \subset \{0, 1, \cdots, n-1\}$) when restricted to $\{-1, 1\}^n$, since $x_k^2 = 1$. Note that we have defined $\prod_{k \in \{\}} x_k = 1$. The terms in the expression of $p(x_0, x_1 \ldots x_{n-1})$ without the leading coefficients are called monomials (i.e., $\prod_{k \in S_i} x_k$ for some i). The set of monomials that can be generated using $x_0, x_1, \ldots, x_{n-1}$ is denoted by M_n. Formally, $\mathrm{M}_n = \{\prod_{k \in S_i} x_k : S_i \subset \{0, 1, \cdots, n-1\}\}$. Thus, $|\mathrm{M}_n| = 2^n$.*

Definition 4. *We define $\psi(p) : \Theta_n \to \{0, 1, \ldots, 2^n\}$ as the number of monomials contained in the polynomial function p. We will also extend the number of monomials function to operate on sets of polynomial functions so that $\psi(Q)$ is the set of nonnegative integers that are the number of monomials contained in the polynomial functions in $Q \subset \Theta_n$.*

Definition 5. *Given a binary classification problem $C = (S_n^+, S_n^-)$, a solution is a function $f(x) : \{-1, 1\}^n \to \Re$ such that $f(x) > 0$ whenever $x \in S^+$ and $f(x) < 0$ whenever $x \in S^-$. Then we say f solves C. When a polynomial function (p) solves a binary classification problem (C), we say that the monomials of p solves C as well as p solves C. Furthermore, the problem C is said to have a solution with $\psi(p)$ number of monomials.*

Definition 6. *Given a binary classification problem $C = (S_n^+, S_n^-)$, a solution set is the collection of polynomial functions that solves C. We define $\Omega_n^{(S^+, S^-)} \subset \Theta_n$ to be the set of all polynomial function solutions to the classification problem (S_n^+, S_n^-). We also use Ω_n^Y when the label Y uniquely identifies the classification problem under consideration.*

Definition 7. *Given a binary classification problem $C = (S_n^+, S_n^-)$, the density of $C = (S_n^+, S_n^-)$ is defined to be the minimum element of $\psi(\Omega_n^{(S^+, S^-)})$.*

Definition 8. *The minimum number of monomials that suffices to separate any dichotomy of $\{-1, 1\}^n$ is defined to be the maximum density associated with $\{-1, 1\}^n$ (or n). Formally we have*

$$\Pi(n) = max \bigcup_{C \in \Lambda_n} min\, \psi(\Omega_n^C). \tag{2.1}$$

The goal of this letter is to advance our understanding of $\Pi(n)$, the maximum over the densities of the dichotomies of $\{-1, 1\}^n$.

3 Polynomial (Spectral) Representation of Dichotomies of $\{-1, 1\}^n$

The polynomial/spectral representation of Boolean functions (i.e., dichotomies of $\{-1, 1\}^n$) and the standard results are covered in Saks (1993)

and Siu, Roychowdhury, and Kailath (1995). Here we will present only the necessary results without proofs. A dichotomy of $\{-1, 1\}^n$ being equivalent to a Boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ can be represented as a vector in $\{-1, 1\}^{2^n}$ by adopting a fixed ordering over the assignment vectors. Moreover, f has a unique representation as the weighted sum of monomials with coefficients $\mathbf{a} = (a_1, a_2, \ldots, a_{2^n})^T \in \Re^{2^n}$, called the *spectral coefficients*,

$$f(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{2^n} a_i \prod_{k \in S_i} x_k \text{ where } S_i \subset \{0, 1, \cdots, n-1\}.$$

Noting that each monomial is also a Boolean function, we switch to vector notation and write $\mathbf{f} = \mathbf{D}^n\mathbf{a}$ where the columns of $\mathbf{D}^n$ are the vector representations of the monomials. With appropriate ordering of the monomials[1] and assignment vectors,[2] $\mathbf{D}^n$ becomes a so-called Sylvester-type Hadamard matrix (Bruck, 1990; Siu et al., 1995), which has the following properties.

Lemma 1. *$\boldsymbol{D}^n$ satisfies the recursive relation*

$$\boldsymbol{D}^1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \boldsymbol{D}^{n+1} = \begin{bmatrix} \boldsymbol{D}^n & \boldsymbol{D}^n \\ \boldsymbol{D}^n & -\boldsymbol{D}^n \end{bmatrix} \text{for } n > 0. \tag{3.1}$$

Lemma 2. *$\boldsymbol{D}^n$ is symmetric.*

Lemma 3. *$\boldsymbol{D}^n\boldsymbol{D}^n = 2^n\boldsymbol{I}$.*

Corollary 1. *The inverse of $\boldsymbol{D}^n$* is $(\boldsymbol{D}^n)^{-1} = 2^{-n}\boldsymbol{D}^n$.

Corollary 2. *The matrix $\hat{\boldsymbol{D}}^n = 2^{-n/2}\boldsymbol{D}^n$ is orthogonal.*

4 The Set of Solving Polynomial Functions

Definition 9 (standard form). *Assume $p \in \Theta_n$ with $p(\mathbf{x}) = a_1 + a_2x_0 + a_3x_1 + a_4x_1x_0 + \cdots + a_{2^n}x_{n-1}x_{n-2}\cdots x_0$ solves a classification problem $C = (S^+, S^-)$. Further assume C is a dichotomy. Let $\boldsymbol{a} = (a_1, a_2, \cdots, a_{2^n})^T \in \Re^{2^n}$. Since $p(S^+) > 0$ and $p(S^-) < 0$, C partitions the rows of $\boldsymbol{D}^n$ into $\boldsymbol{D}^n_+$ and $\boldsymbol{D}^n_-$*

[1] Monomials are ordered as $1, x_0, x_1, x_1x_0, x_2, x_2x_0, x_2x_1, x_2x_1x_0, \ldots, x_{n-1}\ldots x_1x_0$.

[2] Assignments to $(x_0, x_1, x_2, \ldots, x_{n-1})$ are ordered as (0's represent 1's and 1's represent -1): $000\ldots0, 100\ldots0, 010\ldots0, 110\ldots0, 001\ldots0, \ldots, 111\ldots1$.

with $\boldsymbol{D}^n_+\mathbf{a} > \boldsymbol{0}$ *and* $\boldsymbol{D}^n_-\boldsymbol{a} < \boldsymbol{0}$. *By defining*

$$\boldsymbol{Y} = diag([\, y_1 \;\; y_2 \;\cdots\; y_{2^n} \,]) \quad where \quad y_i = \begin{cases} -1 & if \quad i\text{th assignment} \in S^- \\ +1 & if \quad i\text{th assignment} \in S^+ \end{cases}$$

the problem can be written as $\boldsymbol{YD}^n\boldsymbol{a} > \boldsymbol{0}$. *Then a solution* $\boldsymbol{a}$ *to this inequality system provides the coefficients of the polynomial function that is a solution to the classification problem* $C = (S^+, S^-)$. *We call this representation the standard form*.

Assume we are given a problem in the standard form $\mathbf{YD}^n\mathbf{a} > \mathbf{0}$. Then for a solution $\mathbf{a}$, there exists a positive $\mathbf{k} = (k_1 k_2, \ldots, k_{2^n})^T > \mathbf{0}$ such that $\mathbf{YD}^n\mathbf{a} = \mathbf{k}$. Noting that $\mathbf{Y}^{-1} = \mathbf{Y}$ and $(\mathbf{D}^n)^{-1} = 2^{-n}\mathbf{D}^n$ (see corollary 1), we can solve the coefficients of p as $\mathbf{a} = 2^{-n}\mathbf{D}^n\mathbf{Yk}$. Thus, a solution to the problem $C = (S^+, S^-)$ is a positive linear combination of the columns of $\mathbf{D}^n\mathbf{Y}$ (or rows of $\mathbf{YD}^n$). Conversely, assume $\mathbf{b}$ is a positive combination of the columns of $\mathbf{D}^n\mathbf{Y}$ so that $\mathbf{b} = \mathbf{D}^n\mathbf{Yk}$ for some $\mathbf{k} > \mathbf{0}$. Then $\mathbf{YD}^n\mathbf{b} = 2^n\mathbf{k} > \mathbf{0}$, implying that

$$\begin{aligned} q(x_0, x_1 \cdots x_{n-1}) = {} & b_1 + b_2 x_0 + b_3 x_1 \\ & + b_4 x_1 x_0 + \cdots + b_{2^n} x_{n-1} x_n \ldots x_0 \in \Theta_n x \end{aligned}$$

is a solution to $C = (S^+, S^-)$. So we have established the following result:

Theorem 1. *Given a dichotomy of dimension n in the standard form,* $\boldsymbol{YD}^n\boldsymbol{a} > \boldsymbol{0}$, *the set of solutions,* $\Omega_n^Y \subset \Theta_n$ *is exactly the polynomial functions with the coefficients taken from the interior of the cone defined by the rows of* $\boldsymbol{YD}^n$. *In short, we write* $\Omega_n^Y = int\ cone(\boldsymbol{YD}^n)$.

5 Main Theorem: An Upper Bound for the Minimum Number of Monomials

Theorem 2 (main theorem). *For any binary classification problem in* $\{-1, 1\}^n$, $n > 1$, *there exists always a polynomial function solution with* $2^n - 2^n/4$ *or fewer monomials. Equivalently, the maximum density over all the n-dimensional Boolean functions is bounded (from above) by* $2^n - 2^n/4$. *Formally,*

$$max \bigcup_{C \in \Lambda_n} min\ \psi(\Omega_n^C) = \Pi(n) \leq 2^n - 2^n/4. \tag{5.1}$$

Proof. We prove the theorem by constructing a polynomial function solution with $2^n - 2^n/4$ or fewer monomials for an arbitrary dichotomy of $\{-1, 1\}^n$.

Any dichotomy of $\{-1,1\}^n$ is characterized by the standard form $diag(\mathbf{y})\mathbf{D}^n\mathbf{z} > \mathbf{0}$ for some $\mathbf{y} \in \{-1,1\}^{2^n}$. A solution vector $\mathbf{z}$ gives the monomial coefficients of the solving polynomial function. Thus, we have to prove that a solution to the inequality system exists with at least one-fourth of the components of $\mathbf{z}$ equal to zero.

Using lemma 1, write $\mathbf{D}^n$ in terms of $\mathbf{D}^{n-1}$, and partition $\mathbf{y}$ and $\mathbf{z}$ into two halves:

$$diag(\mathbf{y})\mathbf{D}^n\mathbf{z} > \mathbf{0} \Leftrightarrow diag\left(\begin{bmatrix}\mathbf{y}^u\\ \mathbf{y}^d\end{bmatrix}\right)\begin{bmatrix}\mathbf{D}^{n-1} & \mathbf{D}^{n-1}\\ \mathbf{D}^{n-1} & -\mathbf{D}^{n-1}\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ \mathbf{t}\end{bmatrix} > \mathbf{0}. \tag{5.2}$$

Further, write the submatrices $\mathbf{D}^{n-1}$ as row vectors:

$$diag\left(\begin{bmatrix}\mathbf{y}^u\\ \mathbf{y}^d\end{bmatrix}\right)\left[\begin{array}{c|c}\mathbf{d}_1 & \mathbf{d}_1\\ \vdots & \vdots\\ \mathbf{d}_{2^{n-1}} & \mathbf{d}_{2^{n-1}}\\ \hline \mathbf{d}_1 & -\mathbf{d}_1\\ \vdots & \vdots\\ \mathbf{d}_{2^{n-1}} & -\mathbf{d}_{2^{n-1}}\end{array}\right]\begin{bmatrix}\mathbf{w}\\ \mathbf{t}\end{bmatrix} > \mathbf{0}. \tag{5.3}$$

By expanding (5.3) we see that for each $i \in \{1..2^{n-1}\}$ we have

$$y_i^u\,(\mathbf{d_i}\mathbf{w} + \mathbf{d_i}\mathbf{t}) > \mathbf{0} \text{ and } y_i^d\,(\mathbf{d_i}\mathbf{w} - \mathbf{d_i}\mathbf{t}) > \mathbf{0}. \tag{5.4}$$

Depending on the values of the components of $\mathbf{y}$, we have four cases; we first look at the following two:

$$\begin{aligned}(y_i^u, y_i^d) = (+1,+1) &\Rightarrow \mathbf{d_i}\mathbf{w} > \mathbf{d_i}\mathbf{t} > -\mathbf{d_i}\mathbf{w}\\ (y_k^u, y_k^d) = (-1,-1) &\Rightarrow -\mathbf{d_k}\mathbf{w} > \mathbf{d_k}\mathbf{t} > \mathbf{d_k}\mathbf{w}\\ &\Rightarrow (-\mathbf{d_k})\mathbf{w} > (-\mathbf{d_k})\mathbf{t} > -(-\mathbf{d_k})\mathbf{w}.\end{aligned} \tag{5.5}$$

We construct the $r(\mathbf{F}) \times 2^{n-1}$ matrix $\mathbf{F}$ using the rows that satisfy equation 5.5, namely, with $\mathbf{d_i}$ and $-\mathbf{d_k}$. Then equation 5.5 can be compactly expressed as $\mathbf{F}\mathbf{w} > \mathbf{F}\mathbf{t} > -\mathbf{F}\mathbf{w}$. Note we allow $r(\mathbf{F}) = 0$, meaning that $\mathbf{F}$ is the empty matrix (see the Remark below).

Next, consider the remaining two cases:

$$\begin{aligned}(y_i^u, y_i^d) = (+1,-1) &\Rightarrow \mathbf{d_i}\mathbf{t} > \mathbf{d_i}\mathbf{w} > -\mathbf{d_i}\mathbf{t}\\ (y_k^u, y_k^d) = (-1,+1) &\Rightarrow -\mathbf{d_k}\mathbf{t} > \mathbf{d_k}\mathbf{w} > \mathbf{d_k}\mathbf{t}\\ &\Rightarrow (-\mathbf{d_k})\mathbf{t} > (-\mathbf{d_k})\mathbf{w} > -(-\mathbf{d_k})\mathbf{t}.\end{aligned} \tag{5.6}$$

Similarly, we construct the matrix $\mathbf{G}$ of size $r(\mathbf{G}) \times 2^{n-1}$ using the rows that satisfy equation 5.6, namely, with $\mathbf{d_i}$ and $-\mathbf{d_k}$. Then equation 5.6 can be compactly stated as $\mathbf{Gt} > \mathbf{Gw} > -\mathbf{Gt}$. Note we allow $r(\mathbf{G}) = 0$, meaning that $\mathbf{G}$ is the empty matrix (see the Remark below).

A simple but useful identity regarding the size of $\mathbf{F}$ and $\mathbf{G}$ is

$$r(\mathbf{F}) + r(\mathbf{G}) = 2^{n-1}. \tag{5.7}$$

Thus, we have obtained two coupled systems of inequalities in the 2^{n-1} dimension (if $r(\mathbf{F}) = 0$ or $r(\mathbf{G}) = 0$ there will be a single inequality system):

$$\mathbf{Fw} > \mathbf{Ft} > -\mathbf{Fw} \text{ and } \mathbf{Gt} > \mathbf{Gw} > -\mathbf{Gt}. \tag{5.8}$$

Note that the systems are satisfied for all $\mathbf{w} \in \text{int}\, cone(\mathbf{F})$ and $\mathbf{t} \in \text{int}\, cone(\mathbf{G})$ because the rows of $\mathbf{F}$ and $\mathbf{G}$ are mutually orthogonal. Our goal is to find a solution vector $\mathbf{z} = [\mathbf{w}, \mathbf{t}]$ with as many zero components as possible. We now show that equation 5.8 enables us to drive a lower bound for the number of zeros we can obtain in $\mathbf{z} = [\mathbf{w}, \mathbf{t}]$.

Remark. If $r(\mathbf{F}) = 0$, the theorem's claim is readily satisfied by taking $\mathbf{w} = \mathbf{0}$ and $\mathbf{t} \in \text{int}\, cone(\mathbf{G})$. The same is true for $r(\mathbf{G}) = 0$ with the choice of $\mathbf{t} = \mathbf{0}$ and $\mathbf{w} \in \text{int}\, cone(\mathbf{F})$. In this case, the problem is equivalent to a problem in one lower dimension and assumes at least a 2^{n-1} monomial solution.

Now write equation 5.8 in terms of $(\mathbf{w} - \mathbf{t})$ and $(\mathbf{w} + \mathbf{t})$ to get

$$\begin{array}{ll} \mathbf{F}(\mathbf{w} - \mathbf{t}) > \mathbf{0} & \mathbf{F}(\mathbf{w} + \mathbf{t}) > \mathbf{0} \\ -\mathbf{G}(\mathbf{w} - \mathbf{t}) > \mathbf{0} & \mathbf{G}(\mathbf{w} + \mathbf{t}) > \mathbf{0}; \end{array} \tag{5.9}$$

equivalently,

$$\begin{bmatrix} \mathbf{F} \\ -\mathbf{G} \end{bmatrix} (\mathbf{w} - \mathbf{t}) > \mathbf{0} \quad \begin{bmatrix} \mathbf{F} \\ \mathbf{G} \end{bmatrix} (\mathbf{w} + \mathbf{t}) > \mathbf{0}. \tag{5.10}$$

Notice that we have decomposed the original problem into two subproblems of the standard form $diag(\mathbf{y}')\mathbf{D}^{n-1}\mathbf{z}' > \mathbf{0}$ for some $\mathbf{y}'$. Therefore, by theorem 1, the vectors $(\mathbf{w} - \mathbf{t})^T$ and $(\mathbf{w} + \mathbf{t})^T$ must belong to the interior of the cones spanned by the rows of $\{\mathbf{F}, -\mathbf{G}\}$ and $\{\mathbf{F}, \mathbf{G}\}$, respectively. Therefore all the solutions to equation 5.9 are characterized by arbitrary positive real

row vectors $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\gamma}, \boldsymbol{\gamma}'$,

$$\begin{aligned}(\mathbf{w}-\mathbf{t})^T &= 2\boldsymbol{\alpha}F - 2\boldsymbol{\gamma}G \quad \boldsymbol{\alpha}, \boldsymbol{\gamma} > \mathbf{0} \\ (\mathbf{w}+\mathbf{t})^T &= 2\boldsymbol{\alpha}'F + 2\boldsymbol{\gamma}'G \quad \boldsymbol{\alpha}', \boldsymbol{\gamma}' > \mathbf{0},\end{aligned} \tag{5.11}$$

which easily yield expressions for $\mathbf{w}$ and $\mathbf{t}$:

$$\left.\begin{aligned}\mathbf{w}^T &= (\boldsymbol{\alpha}+\boldsymbol{\alpha}')\mathbf{F} + (-\boldsymbol{\gamma}+\boldsymbol{\gamma}')\mathbf{G} \\ \mathbf{t}^T &= (-\boldsymbol{\alpha}+\boldsymbol{\alpha}')\mathbf{F} + (\boldsymbol{\gamma}+\boldsymbol{\gamma}')\mathbf{G}\end{aligned}\right\} \quad \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}. \tag{5.12}$$

Now either $r(\mathbf{G}) \geq r(\mathbf{F})$ or $r(\mathbf{F}) > r(\mathbf{G})$. Assume the former and choose $\boldsymbol{\alpha}, \boldsymbol{\alpha}' > \mathbf{0}$ arbitrarily to get

$$\mathbf{w}^T = (w_1\ w_2\ w_3,\ \ldots,\ w_{2^{n-1}}) + (-\boldsymbol{\gamma}+\boldsymbol{\gamma}')\mathbf{G} \qquad \boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}\,. \tag{5.13}$$

Now consider $\tilde{\mathbf{G}}$, the reduced row-echelon form of $\mathbf{G}$: since the rows of $\mathbf{G}$ are orthogonal, $\tilde{\mathbf{G}}$ has no zero rows. So we can define i_c to be the column index of the leading nonzero element of $\tilde{\mathbf{G}}_i$, the ith row of $\tilde{\mathbf{G}}$. Consider the row vector

$$\mathbf{v} = \sum_{i=1}^{r(\mathbf{G})} -w_{i_c}\tilde{\mathbf{G}}_i. \tag{5.14}$$

Clearly $\mathbf{v}$ is a linear combination of the rows of $\mathbf{G}$, that is,

$$\exists \boldsymbol{\beta} \in \Re^{r(\mathbf{G})} \text{ such that } \boldsymbol{\beta}G = \mathbf{v}. \tag{5.15}$$

Since $\mathbf{G}\mathbf{G}^T = 2^{n-1}\mathbf{I}_{r(\mathbf{G})}$, $\boldsymbol{\beta}$ can be found by the projection (and scaling) of $\mathbf{v}$ onto the rows of $\mathbf{G}$: $\boldsymbol{\beta} = 2^{-(n-1)}\mathbf{v}\mathbf{G}^T$. As $\boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}$ are free parameters in equation 5.13, we can choose them to construct $\boldsymbol{\beta}$ (and hence $\mathbf{v}$). Formally stated,

$$\forall \boldsymbol{\beta} \in \Re^{r(G)}, \exists \boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \Re^{r(G)} > \mathbf{0} \text{ such that } (-\boldsymbol{\gamma}+\boldsymbol{\gamma}') = \boldsymbol{\beta}. \tag{5.16}$$

But because of the construction of $\mathbf{v}$ given in equation 5.14, $\mathbf{w}^T = (w_0\ w_1\ w_2\ \cdots\ w_{2^{n-1}-1}) + \mathbf{v}$ must have at least $r(\mathbf{G})$ zeros.

The value of $r(\mathbf{G})$ can easily be bounded from below. Since $r(\mathbf{G}) \geq r(\mathbf{F})$ and $r(\mathbf{G}) + r(\mathbf{F}) = 2^{n-1}$, we have

$$r(\mathbf{G}) \geq 2^{n-1} - r(\mathbf{G}) \Rightarrow r(\mathbf{G}) \geq 2^n/4. \tag{5.17}$$

Table 2: Example Problem Used in Section 6 and the Verification of the Solution Found by Application of the Main Theorem.

Example Problem				The Solution via Theorem 1	
$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	Class	Sign of $p(x_0, x_1, x_2)$	$p(x_0, x_1, x_2)$
1	1	1	−1	−1	−8
−1	1	1	−1	−1	−4
1	−1	1	1	1	8
−1	−1	1	1	1	16
1	1	−1	1	1	16
−1	1	−1	−1	−1	−4
1	−1	−1	−1	−1	−16
−1	−1	−1	−1	−1	−8

Thus, the solution $\mathbf{z} = [\mathbf{w}, \mathbf{t}]$ has $2^n - 2^n/4$ or fewer nonzero elements, so the corresponding polynomial function has $2^n - 2^n/4$ or fewer monomials.

The case for $r(\mathbf{F}) > r(\mathbf{G})$ is proven in the same way by changing the roles of $\mathbf{F}$, $(\boldsymbol{\alpha}, \boldsymbol{\alpha}')$, $\mathbf{w}$ with $\mathbf{G}$, $(\boldsymbol{\gamma}, \boldsymbol{\gamma}')$, $\mathbf{t}$, respectively. Combining the proofs for the two cases, we conclude that the number of zeros is at least $\max_{\mathbf{F},\mathbf{G}}(r(\mathbf{F}), r(\mathbf{G})) = 2^n/4$. Thus, for any binary classification problem in $\{-1, 1\}^n$, there exists a polynomial function solution with $2^n - \max_{\mathbf{F},\mathbf{G}}(r(\mathbf{F}), r(\mathbf{G})) = 2^n - 2^n/4$ or fewer monomials.

6 An Example

We apply the theorem to the dichotomy of $\{-1, 1\}^3$ given in Table 2 (consider the left four columns). We write the problem in the standard form $diag(\mathbf{y})\mathbf{D}^3\mathbf{z} > \mathbf{0}$, where $\mathbf{y} = (-1, -1, 1, 1, 1, -1, -1, -1)$ is formed by copying the class labels in the order they appear in the table. Applying the partitioning used in the theorem, we have

$$\mathbf{y}^u = (-1, -1, 1, 1), \quad \mathbf{y}^d = (1, -1, -1, -1), \quad \mathbf{z} = [\mathbf{w}, \mathbf{t}] \text{ and}$$

$$\mathbf{D}^3 = \begin{bmatrix} \mathbf{D}^2 & \mathbf{D}^2 \\ \mathbf{D}^2 & -\mathbf{D}^2 \end{bmatrix} \text{ where } \mathbf{D}^2 = \begin{bmatrix} \mathbf{d_1} \\ \mathbf{d}_2 \\ \mathbf{d}_3 \\ \mathbf{d}_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Considering components $\mathbf{y}^u$ and $\mathbf{y}^d$, $(-1, 1)$, $(-1, -1)$, $(1, -1)$, $(1, -1)$, we construct $\mathbf{F}$ and $\mathbf{G}$ matrices as instructed in the theorem, finding

$$\mathbf{F} = \begin{bmatrix} -1 & 1 & -1 & 1 \end{bmatrix} \text{ and } \mathbf{G} = \begin{bmatrix} -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

The number of rows in $\mathbf{F}$ and $\mathbf{G}$ are 1 and 3, respectively, so $r(\mathbf{F}) = 1$ and $r(\mathbf{G}) = 3$. Since $r(\mathbf{G}) \geq r(\mathbf{F})$, we work on the $\mathbf{w}$ part of the solution vector. In the expression for $\mathbf{w}^T = (\boldsymbol{\alpha} + \boldsymbol{\alpha}')\mathbf{F} + (-\boldsymbol{\gamma} + \boldsymbol{\gamma}')\mathbf{G}$, we can choose $\boldsymbol{\alpha}, \boldsymbol{\alpha}' > \mathbf{0}$ arbitrarily. Let us set $\boldsymbol{\alpha} = \boldsymbol{\alpha}' = 0.5$ (note that since $r(\mathbf{F}) = 1$, $\boldsymbol{\alpha}, \boldsymbol{\alpha}'$ became scalars) to have

$$\mathbf{w}^T = (-1 \quad 1 \quad -1 \quad 1) + (-\boldsymbol{\gamma} + \boldsymbol{\gamma}')\mathbf{G}.$$

Applying row echelon reduction to $\mathbf{G}$ yields

$$\tilde{\mathbf{G}} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

So $1_c = 1$, $2_c = 2$, $3_c = 3$. Applying the formula $\mathbf{v} = \sum_{i=1}^{r(\mathbf{G})} -w_{i_c}\tilde{\mathbf{G}}_i$ we obtain

$$\mathbf{v} = -(-1)\tilde{\mathbf{G}}_1 - (1)\tilde{\mathbf{G}}_2 - (-1)\tilde{\mathbf{G}}_3 \quad \Rightarrow \mathbf{v} = (1 \quad -1 \quad 1 \quad 3).$$

Plugging $\mathbf{v}$ in $\boldsymbol{\beta} = 2^{-(n-1)}\mathbf{v}\mathbf{G}^T$, we get

$$\boldsymbol{\beta} = 2^{-2}(1\ -1\ 1\ 3)\mathbf{G}^T \Rightarrow \boldsymbol{\beta} = (-1\ -1\ 1).$$

Now we have infinitely many ways of choosing positive vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ to satisfy $(-\boldsymbol{\gamma} + \boldsymbol{\gamma}') = \boldsymbol{\beta}$. Let us choose $\boldsymbol{\gamma} = (2\ 2\ 1)$ and $\boldsymbol{\gamma}' = (1\ 1\ 2)$. According to the theorem, $\mathbf{w}^T = (-1\ 1\ -1\ 1) + (-\boldsymbol{\gamma} + \boldsymbol{\gamma}')\mathbf{G}$ must have at least $r(\mathbf{G}) = 3$ zeros and $\mathbf{z} = [\mathbf{w}, \mathbf{t}]$ must be a solution to the original problem, where $\mathbf{t}$ is given by $\mathbf{t}^T = (-\boldsymbol{\alpha} + \boldsymbol{\alpha}')\mathbf{F} + (\boldsymbol{\gamma} + \boldsymbol{\gamma}')\mathbf{G} = (\boldsymbol{\gamma} + \boldsymbol{\gamma}')\mathbf{G}$. Carrying out the arithmetic, we indeed find $\mathbf{z} = (0\ 0\ 0\ 4\ 3\ -3\ -9\ -3)^T$. Thus, the solving polynomial function is $p(x_0, x_1, x_2) = -3x_2x_1x_0 - 9x_2x_1 - 3x_2x_0 + 3x_2 + 4x_1x_0$.

It is easily verified that the sign of $p(x_0, x_1, x_2)$ satisfies the assignment table (compare class labels with the last two columns of Table 2). Figure 1 shows the separation obtained by this solution. The surface shown is the contour plot of $p(x_0, x_1, x_2)$ at 0.

This example demonstrates the freedom of choice in constructing a solution. One suspects that a better bound can be obtained by studying the freedom provided by $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}$. In fact, we will prove in section 9 that all dichotomies of $\{-1, 1\}^3$ can be solved with four monomials, which is fewer than the five-monomial solution found by the application of the main theorem to the example problem.

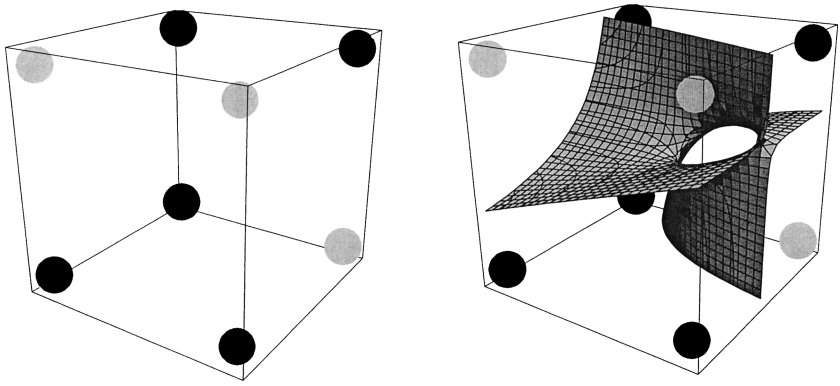

Figure 1: (Left) Depiction of the problem given in section 6. (Right) Solution obtained by the application of the main theorem. The surface has two sides, each facing exclusively the circles from a single class.

7 Fourier-Motzkin Elimination for Binary Classification Problems

Fourier-Motzkin (FM) elimination is a method for eliminating variables from a system of inequalities. It is often used to determine the solvability of the system and finding a feasible solution if it is solvable. Here we will show that FM elimination can be used to construct polynomial function solutions (with fewer monomials) for binary classification problems. First, we introduce the elimination procedure.

7.1 Fourier-Motzkin Elimination. Given an inequality system, the aim of FM elimination is to produce a new inequality system with fewer variables. The key step in the procedure is the elimination of a single variable, which is repeatedly applied to the current inequality system until no variable can be eliminated. If the elimination yields an inconsistent inequality system, then it is concluded that the original system has no solution. Here we assume that the inequality system is given by $\mathbf{Ax} > \mathbf{0}$. For a more general treatment readers are referred to other sources (e.g., Chandru, 1993).

Suppose we wish to eliminate the variable x_j (or column j) from $\mathbf{Ax} > \mathbf{0}$. Define

$$\begin{aligned} \mathrm{I}^+ &= \{i : \mathbf{A}_{ij} > 0\} \\ \mathrm{I}^- &= \{i : \mathbf{A}_{ij} < 0\} \\ \mathrm{I}^0 &= \{i : \mathbf{A}_{ij} = 0\}. \end{aligned} \tag{7.1}$$

If $I^- = \{\}$ or $I^+ = \{\}$, then x_j cannot be eliminated. Assume this is not the case. We create a new matrix $\mathbf{A}'$ with the row entries taken from the set $\{|A_{kj}|\mathbf{A}_i + |A_{ij}|\mathbf{A}_k : i \in I^-, k \in I^+\} \cup I^0$. The new inequality system $\mathbf{A}'\mathbf{x} > \mathbf{0}$ has zero coefficients for x_j. Thus, we can write a reduced system $\tilde{\mathbf{A}}\tilde{x} > \mathbf{0}$ (by removing column j from $\mathbf{A}'$ and x_j from $\mathbf{x}$). Clearly a solution to the system $\mathbf{Ax} > \mathbf{0}$ is a solution to the system $\tilde{\mathbf{A}}\tilde{x} > \mathbf{0}$ since it is constructed with elementary row operations that involve positive scaling and addition of the rows of $\mathbf{A}$. The converse is also true, as stated next (for a proof, see Chandru, 1993).

Proposition 1. *Given the inequality system $\boldsymbol{Ax} > \boldsymbol{0}$, consider the one-variable (say, x_1) eliminated system $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} > \boldsymbol{0}$. Then for all the solutions of the reduced system, it is guaranteed that there will be a value for x_1 such that the original inequality system will be satisfied with $\boldsymbol{x} = [x_1, \tilde{\boldsymbol{x}}]$.*

7.2 Polynomial Function Solution via Fourier-Motzkin Elimination. Given a classification problem in the standard form $\mathbf{YD}^n\mathbf{a} > \mathbf{0}$, the idea is to pick an elimination order and eliminate the matrix $\mathbf{YD}^n$ regarding $\mathbf{a}$ as the vector of variables. The resultant matrix then can be easily converted into a solution vector with the zero components corresponding to the eliminated columns of $\mathbf{YD}^n$.

Definition 10. *Given an inequality system $\boldsymbol{Ax} > \boldsymbol{0}$, we use $\boldsymbol{A}^\circ$ to denote the matrix after the repeated application of FM elimination to all the columns of $\boldsymbol{A}$. The order of elimination is indeterminate, so $\boldsymbol{A}^\circ$ is in general ambiguously defined. When no order is specified, an arbitrary order is implied. Note that $\boldsymbol{A}^\circ$ has the same number of columns as $\boldsymbol{A}$ but with zero columns corresponding to the eliminated variables.*

Proposition 2. *Assume that we are given a classification problem in the standard form $\boldsymbol{YD}^n\boldsymbol{a} > \boldsymbol{0}$. Let $\boldsymbol{Q} = \boldsymbol{YD}^n$ so that we have the system of inequalities $\boldsymbol{Qa} > \boldsymbol{0}$. Apply FM elimination to all columns of $\boldsymbol{Q}$ to obtain $\boldsymbol{Q}^\circ\boldsymbol{a} > \boldsymbol{0}$. Then the row sum given by $\boldsymbol{c} = \sum_{i=1}^m \boldsymbol{Q}_i^\circ$ is a solution, that is, $\boldsymbol{YD}^n\boldsymbol{c}^T > \boldsymbol{0}$, where m is the number of rows of $\boldsymbol{Q}^\circ$.*

Proof. Clearly $\mathbf{c}^T \in \text{int}\, cone(\mathbf{YD}^n)$. Therefore due to theorem 1, $\mathbf{c}$ must satisfy $\mathbf{YD}^n\mathbf{c}^T > \mathbf{0}$. In fact, for any $w_i > 0$, the sum $\mathbf{c} = \sum_{i=1}^m w_i\mathbf{Q}_i^\circ$ is also a solution.

Definition 11. *We call the row vector $\boldsymbol{c} = \sum_{i=1}^m \boldsymbol{Q}_i^\circ$ the FM sum.*

7.3 Example: Polynomial Function Solution via FM Elimination. Consider the two-dimensional classification problem specified in Table 3.

Table 3: Assignment Table for the Two-Dimensional Example Problem.

x_0	x_1	Class
1	1	−1
−1	1	1
1	−1	1
−1	−1	−1

We write the problem in the standard form $\mathbf{YD^2a} > \mathbf{0}$:

$$\underbrace{\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}}_{\mathbf{YD^2a}} = \underbrace{\begin{bmatrix} -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}}_{\mathbf{Qa}} > \mathbf{0}.$$

Let us eliminate column 1 from $\mathbf{Q}$. Since $I^+ = \{2, 3\}$, $I^- = \{1, 4\}$ and $I^0 = \{\}$, we get

$$\mathbf{Q}_1 = \begin{bmatrix} 0 & -2 & 0 & -2 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & -2 & -2 \\ 0 & 2 & 0 & -2 \end{bmatrix}.$$

Eliminate column 2 from $\mathbf{Q}_1$: $I^+ = \{4\}$, $I^- = \{1\}$, $I^0 = \{2, 3\}$, so we have

$$\mathbf{Q}_2 = \begin{bmatrix} 0 & 0 & 2 & -2 \\ 0 & 0 & -2 & -2 \\ 0 & 0 & 0 & -8 \end{bmatrix}.$$

Eliminate column 3 from $\mathbf{Q}_2$: $I^+ = \{1\}$, $I^- = \{2\}$,$I^0 = \{3\}$, so we have

$$\mathbf{Q}_3 = \begin{bmatrix} 0 & 0 & 0 & -8 \\ 0 & 0 & 0 & -8 \end{bmatrix}.$$

We cannot eliminate any more columns, so $\mathbf{Q}^\circ = \mathbf{Q}_3$; thus, the sum of the rows of $\mathbf{Q}^\circ$ must be the coefficients of a solution. Namely, we have $p(x_0, x_1) = -16x_1x_0$. In this case, FM elimination has found the minimum number of monomial solutions to the problem.

8 Extension of the Main Theorem

Although the extension theorem gives a slight improvement, it shows how one might pursue a proof based on theorem 2 and FM elimination to improve the bound. The reader might have already noticed the freedom of parameter choice in theorem 2, which suggests that the bound can indeed be improved.

Theorem 3 (extension of the main theorem). *For any binary classification problem in $\{-1, 1\}^n$, $n > 2$, there exists always a polynomial function solution with $2^n - 2^n/4 - 1$ or fewer monomials. Formally stated,*

$$max \bigcup_{C \in \Lambda_n} min\psi(\Omega_n^C) = \Pi(n) \leq 2^n - 2^n/4 - 1. \tag{8.1}$$

First we prove two lemmas.

Lemma 4. *Given an arbitrary (row) vector $\boldsymbol{\beta}$ and a positive (row) vector $\boldsymbol{\delta}^q > \mathbf{0}$, the system of equations with the unknown vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$,*

$$\begin{array}{l}(-\boldsymbol{\gamma} + \boldsymbol{\gamma}') = \boldsymbol{\beta} \\ (\boldsymbol{\gamma} + \boldsymbol{\gamma}') = M\boldsymbol{\delta}^q\end{array} \text{ with } M = (1+\varepsilon)\frac{max(|\boldsymbol{\beta}_j|)}{min(\boldsymbol{\delta}_k^q)}, \varepsilon > 0, \tag{8.2}$$

where max and min runs over the components of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}^q$, and $\varepsilon > 0$ always has a positive solution $\boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}$.

Proof. By solving the system for $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$, we see the solutions have to be positive because of the construction of M:

$$\begin{aligned}\gamma'_i &= 0.5(M\delta_i^q + \beta_i) = 0.5\left((1+\varepsilon)\frac{\max(|\boldsymbol{\beta}_j|)}{\min(\boldsymbol{\delta}_k^q)}\delta_i^q + \beta_i\right) \\ &\geq 0.5\left((1+\varepsilon)\max(|\boldsymbol{\beta}_j|) + \beta_i\right) > 0 \\ \gamma_i &= 0.5(M\delta_i^q - \beta_i) = 0.5\left((1+\varepsilon)\frac{\max(|\boldsymbol{\beta}_j|)}{\min(\boldsymbol{\delta}_k^q)}\delta_i^q - \beta_i\right) \\ &\geq 0.5\left((1+\varepsilon)\max(|\boldsymbol{\beta}_j|) - \beta_i\right) > 0.\end{aligned} \tag{8.3}$$

Thus, $\boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}$ as desired. Note that that $M\boldsymbol{\delta}^q\mathbf{G}$ and $\boldsymbol{\delta}^q\mathbf{G}$ will have the same number of zero components ($\mathbf{G}$ is an appropriately sized real matrix).

Lemma 5. *Assume we have constructed $\boldsymbol{F}$ and $\boldsymbol{G}$ matrices as in theorem 2 and $r(\boldsymbol{G}) \geq r(\boldsymbol{F})$ for a given problem. Then assume that there exists a positive row vector $\boldsymbol{\delta}^q > \mathbf{0}$ such that $\boldsymbol{\delta}^q\boldsymbol{G}$ has q zero components. Furthermore, assume there exist*

row vectors $\boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}$ *satisfying both* $(-\boldsymbol{\gamma} + \boldsymbol{\gamma}') = \boldsymbol{\beta}$($\boldsymbol{\beta}$ *as defined in theorem 2) and* $(\boldsymbol{\gamma} + \boldsymbol{\gamma}') = \boldsymbol{\delta}^q$. *Then the number of zeros in the solution can be improved to* $2^{n-2} + q$.

Proof. We are given $r(\mathbf{G}) \geq r(\mathbf{F})$. Choose $\boldsymbol{\alpha} = \boldsymbol{\alpha}' = 0.5\mathbf{I}_{r(\mathbf{G})}$. From theorem 2, we know that if we choose $\boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}$ such that $(-\boldsymbol{\gamma} + \boldsymbol{\gamma}') = \boldsymbol{\beta}$, the half solution vector $\mathbf{w}$ given by $\mathbf{w}^T = (w_0\ w_1\ w_2, \ldots, w_{2^{n-1}-1}) + (-\boldsymbol{\gamma} + \boldsymbol{\gamma}')\mathbf{G}$ is guaranteed to have at least 2^{n-2} zeros. The expression for the other half of the solution given by $\mathbf{t}^T = (-\boldsymbol{\alpha} + \boldsymbol{\alpha}')\mathbf{F} + (\boldsymbol{\gamma} + \boldsymbol{\gamma}')\mathbf{G}$ is $\mathbf{t}^T = \boldsymbol{\delta}^q\mathbf{G}$, since $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ and $\boldsymbol{\delta}^q = (\boldsymbol{\gamma} + \boldsymbol{\gamma}')$ by the premises of the lemma. So, the proof is complete because the solution is the concatenation of $\mathbf{w}$ and $\mathbf{t}$.

Proof of Theorem 3. Proceed as in theorem 2 to construct $\mathbf{F}$ and $\mathbf{G}$ matrices. Then either $r(\mathbf{G}) \geq r(\mathbf{F})$ or $r(\mathbf{F}) > r(\mathbf{G})$; assume the former. We are given $n > 2$, so $r(\mathbf{G}) \geq r(\mathbf{F})$ implies $r(\mathbf{G}) > 1$ due to the identity $r(\mathbf{G}) + r(\mathbf{F}) = 2^{n-1}$. Since the rows of $\mathbf{G}$ are orthogonal and taken from $\{-1, 1\}^{n-1}$, there must be a column where not all the components have the same sign. Application of FM elimination on this column and taking FM sum results in a vector with at least one zero component. This means that there exists $\boldsymbol{\delta}' > \mathbf{0}$ such that $\boldsymbol{\delta}'\mathbf{G}$ has at least one zero component. By lemma 4, $\boldsymbol{\delta}'$ can be positively scaled as $\mathbf{q}^1 = M\mathbf{q}'$ such that $(\boldsymbol{\gamma} + \boldsymbol{\gamma}') = \mathbf{q}^1$ and $(-\boldsymbol{\gamma} + \boldsymbol{\gamma}') = \boldsymbol{\beta}$ have a positive solution $\boldsymbol{\gamma}, \boldsymbol{\gamma}' > \mathbf{0}$, and $\mathbf{q}^1\mathbf{G}$ has one zero component. Due to lemma 5, this implies that the number of zeros in the solution can be made at least $2^{n-2} + 1$, proving the theorem.

Remark. The case for $r(\mathbf{F}) > r(\mathbf{G})$ is proven in the same way by changing the roles of $\mathbf{F}$,$(\boldsymbol{\alpha}, \boldsymbol{\alpha}')$, $\mathbf{w}$ with $\mathbf{G}$, $(\boldsymbol{\gamma}, \boldsymbol{\gamma}')$, $\mathbf{t}$, respectively, in lemma 5 and the proof.

9 Some Results on Lower Dimensions

This section presents exact results concerning the minimum number of monomials required to solve the binary classification problems in $\{-1, 1\}^1$, $\{-1, 1\}^2$, and $\{-1, 1\}^3$.

Corollary 3. (corollary to theorem 3). *Any binary classification in* $\{-1, 1\}^3$ *can be solved with four monomials.*

Proof. Proceed as in theorem 2 to construct $\mathbf{F}$ and $\mathbf{G}$ matrices. Then either $r(\mathbf{G}) \geq r(\mathbf{F})$ or $r(\mathbf{F}) > r(\mathbf{G})$; assume the former. Since $r(\mathbf{G}) + r(\mathbf{F}) = 2^2$, we have three cases:

Case 1: $r(\mathbf{G}) = 4, r(\mathbf{F}) = 0$. Due to the remark for theorem 2, there exists a four-monomial solution.

Case 2: $r(\mathbf{G}) = 3, r(\mathbf{F}) = 1$. Application of theorem 2 yields three zeros on the $\mathbf{w}$ part of the solution vector. Following the steps of theorem 3, it is apparent that this solution can be improved by at least one.

Case 3: $r(\mathbf{G}) = 2, r(\mathbf{F}) = 2$. Application of theorem 2 yields two zeros on the $\mathbf{w}$ part of the solution vector. Let $\boldsymbol{\delta}^q = (1, 1)$. Then $\boldsymbol{\delta}^q\mathbf{G}$ must have exactly two zeros since the rows are orthogonal vectors from $\{-1, 1\}^4$. Therefore, due to lemmas 4 and 5, there exists a four-monomial solution.

The case for $r(\mathbf{F}) > r(\mathbf{G})$ is proven similarly.

Proposition 3. *There is a dichotomy in* $\{-1, 1\}^3$ *that cannot be solved with three (or fewer) monomials.*

Proof. We find a problem that cannot be solved with three monomials. The example problem given in section 7 serves the purpose. If the problem has a three-monomial solution, then the inequality system $[\,\mathbf{c}_{j_1}\ \mathbf{c}_{j_2}\ \mathbf{c}_{j_3}\,]\mathbf{a} = \mathbf{Ha} > \mathbf{0}$ must be satisfiable for some $\mathbf{c}_{j_1}$, $\mathbf{c}_{j_2}$, and $\mathbf{c}_{j_3}$, each of them distinct columns of $\mathbf{D}^4$. Satisfiability of $\mathbf{Ha} > \mathbf{0}$ can be checked using FM elimination: if the eliminated system is inconsistent, then $\mathbf{Ha} > \mathbf{0}$ cannot be satisfied due to proposition 1. By applying this procedure for all the $8!/8!\,(8-3)! = 56$ possible cases, it can be shown that that there is no $(j_1,\ j_2,\ j_3)$ that leads to a consistent set of inequalities. Thus, there is no three-monomial solution to the given problem.

Proposition 4. *There is a dichotomy in* $\{-1, 1\}^2$ *that cannot be solved with two or fewer monomials.*

Proof. We find an example problem. Take the classification problem $= (\{(1\ 1), (-1\ 1), (1\ -1)\}, \{(-1\ -1)\})$. Following the logic described in the proof of proposition 3, it can be shown that C cannot be solved with two monomials.

Corollary 4.

i. *Clearly problems in* $\{-1, 1\}^1$ *always require one monomial (one of* x_0 *or 1) solution. Therefore,* $\Pi(1) = 1$.
ii. *The problems in* $\{-1, 1\}^2$ *can always be solved with three monomials (main theorem) and according to proposition 4, there is at least a problem that cannot be solved with fewer than three monomials. Therefore,* $\Pi(2) = 3$.
iii. *Combining corollary 3 (there is always a four-monomial solution) and proposition 3 (there is a dichotomy that cannot be solved with three or fewer monomials), we get* $\Pi(3) = 4$.

Thus, we have established the exact results for the maximum density of the dichotomies of dimensions 1, 2, and 3:

$$\begin{aligned}\Pi(1) &= 1\\ \Pi(2) &= 3\\ \Pi(3) &= 4.\end{aligned} \tag{9.1}$$

It can be shown that $\Pi(4) \leq 9$ (proven with random search) and in fact it appears that $\Pi(4) = 9$ (not proven—empirical observation), tempting one to speculate on the possibility of the general formula,

$$\Pi(n) = \begin{cases} 2^{n-1} & \text{if } n \text{ is odd} \\ 2^{n-1}+1 & \text{if } n \text{ is even} \end{cases}$$

that conforms the known bounds for $n > 1$, that is, $0.11 \times 2^n < \Pi(n) \leq 0.75 \times 2^n$.

10 Conclusion

This letter presented theoretical results regarding the maximum density, $\Pi(n)$, defined as the minimum number of monomials with which one can separate any dichotomy of $\{-1, 1\}^n$. The best-known bound prior to this work was asymptotic and substantially inferior to the proven bound. It is shown that for dimensions 1, 2, and 3, $\Pi(n)$ is equal to 1, 3, and 4, respectively, and less than $2^n - 2^n/4$ for $n > 3$. This result says that given any dichotomy of $\{-1, 1\}^n$, it is always possible to perform the target separation with less than three-quarters of the full set of n-dimensional monomials (2^n monomials). This is the first time a ratio bound independent of n, namely, 3/4, is shown for the maximum density.

In general, a higher-order neuron (HON) would require an exponentially growing number of input lines (number of monomials) to implement a given dichotomy (fully specified binary classification problems). Although this seems to reduce the validity of the HON models of real neurons, one also has to consider that the number of dichotomies grows superexponentially with n. This suggests the possibility that a useful subset of dichotomies might be implemented by HONs with a subexponential number of monomials. Although it is trivially shown that all the classification problems specified at, say, a polynomial number of assignments ($p(n)$) are always solvable with $p(n)$ monomials, the conditions at which a superpolynomial (e.g., $\varepsilon 2^n$ for some, $0 < \varepsilon < 1$) number of assignment specifications would assume a solution with polynomial (i.e., $q(n)$) number of monomials is unexplored. New techniques that use the unspecified assignments to reduce the number of monomials that would suffice to solve a partially specified

problem must be developed, for which this study might provide a starting point. In spite of the success of spectral theory of Boolean functions in obtaining insights on the HON solutions of binary classification problems, it appears that it has certain limitations when the underlying local structure of the Boolean functions (i.e., individual vector components) has to be considered, as is the case when the unspecified assignments need to be exploited for arriving at reduced number of monomial solutions. This is probably why the previous bounds obtained using techniques from the spectral analysis of Boolean functions are inferior to the bound derived in this study, which employs simple local algebraic manipulations.

Acknowledgments

This study was supported by JST-ICORP Computational Brain Project. I was introduced to the problem of establishing a bound on the maximum density by Marifi Guler. I thank Junmei Zhu and Jun Nakanishi for their comments on an earlier version of the manuscript. I thank Mitsuo Kawato, Gordon Cheng, and Hiroshi Imamizu for providing the research environment. Finally I thank an anonymous reviewer for pointing me to the literature on spectral theory of Boolean functions.

References

Bruck, J. (1990). Harmonic analysis of polynomial threshold functions. *SIAM Journal of Discrete Mathematics, 3*, 168–177.

Chandru, V. (1993). Variable elimination in linear constraints. *Computer Journal, 36*, 463–470.

Ghosh, J., & Shin, Y. (1992). Efficient higher order neural networks for classification and function approximation. *International Journal of Neural Systems, 3*, 323–350.

Giles, C. L., & Maxwell, T. (1987). Learning, invariance, and generalization in high-order neural networks. *Applied Optics, 26*, 4972–4978.

Gotsman, C. (1989). *On Boolean functions, polynomials and algebraic threshold functions*. (Tech. Rep. TR-89-18). Tal Aviv: Department of Computer Science, Hebrew University.

Guler, M. (2001). A model with an intrinsic property of learning higher order correlations. *Neural Networks, 14*, 495–504.

Mel, B. W. (1994). Information processing in dendritic trees. *Neural Computation, 6*, 1031–1085.

Mel, B. W., & Koch, C. (1990). Sigma-pi learning: On radial basis functions and cortical associative learning In D. Touretzky (Ed.), *Advances in neural information processing systems, 2* (pp. 474–481). San Mateo, Morgan Kaufmann.

O'Donnell, R., & Servedio, R. (2003). Extremal properties of polynomial threshold functions. In *Eighteenth Annual Conference on Computational Complexity* (pp. 3–12). Piscataway, NJ: IEEE Computer Society.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, the PDP Research

Group (Eds.), Parallel distributed processing (Vol. 1, pp. 151–193). Cambridge, MA: MIT Press.
Saks, M. E. (1993). Slicing the hypercube. In K. Walker (Ed.), *Surveys in combinatorics* (pp. 211–255). Cambridge: Cambridge University Press.
Schmitt, M. (2002). On the complexity of computing and learning with multiplicative neural networks. *Neural Computation, 14*, 241–301.
Schmitt, M. (2005). On the capabilities of higher-order neurons: A radial basis function approach. *Neural Computation, 17*, 715–729.
Siu, K. Y., Roychowdhury, V., & Kailath, T. (1995). Discrete neural computation. Englewood Cliffs, NJ: Prentice Hall.

Received October 11, 2005; accepted May 5, 2006.