

On Specifying Boolean Functions by Labelled Examples

Martin Anthony and Graham Brightwell
Department of Statistical and Mathematical Sciences
London School of Economics
(University of London)
Houghton Street, London WC2A 2AE, U.K.
anthony@vax.lse.ac.uk, brightwe@vax.lse.ac.uk

and

John Shawe-Taylor
Department of Computer Science
Royal Holloway and Bedford New College
(University of London)
Egham Hill, Egham, Surrey TW20 0EX, U.K.
john@cs.rhbnc.ac.uk

Abstract

We say a function t in a set H of $\{0, 1\}$ -valued functions defined on a set X is *specified* by $S \subseteq X$ if the only function in H which agrees with t on S is t itself. The *specification number* of t is the least cardinality of such an S . For a general finite class of functions, we show that the specification number of any function in the class is at least equal to a parameter from [21] known as the testing dimension of the class. We investigate in some detail the specification numbers of functions in the set of linearly separable Boolean functions of n variables—those functions f such that $f^{-1}(\{0\})$ and $f^{-1}(\{1\})$ can be separated by a hyperplane. We present general methods for finding upper bounds on these specification numbers and we characterise those functions which have largest specification number. We obtain a general lower bound on the specification number and we show that for all *nested* functions, this lower bound is attained. We give a simple proof of the fact that for any linearly separable Boolean function, there is exactly one set of examples of minimal cardinality which specifies the function. We discuss those functions which have limited dependence, in the sense that some of the variables are redundant (that is, there are irrelevant attributes), giving tight upper and lower bounds on the specification numbers of such functions. We then bound the average, or expected, number of examples needed to specify a linearly separable Boolean function. In the final section of the paper, we address the complexity of computing specification numbers and related parameters.

1 Introduction

Recent work [11, 21, 22, 23, 15, 24] in computational learning theory has discussed learning in situations where the teacher is helpful and can choose to present carefully chosen sequences of labelled examples to the learner. In this paper we discuss this framework and investigate the number of cleverly-chosen examples needed. Our main contribution is a fairly detailed analysis, using combinatorial and geometrical techniques, of the number of examples required for the set of linearly separable Boolean functions.

1.1 Definitions

A *hypothesis space* H on an *example space* X is a set of $\{0, 1\}$ -valued functions defined on the set X . The individual members of H are called *hypotheses* or *concepts*. We use the standard framework of learning from examples, in which there is some hypothesis t in H , known as the *target*, which is used to classify a sequence of examples presented to a *learner*. More formally, a *sample of length* m is a sequence $\mathbf{x} = (x_1, x_2, \dots, x_m)$ of examples, and the corresponding *training sample* $\mathbf{x}(t)$ for t is this sample, together with the values of t on the examples. We say that x is a *positive* example of t if $t(x) = 1$ and that it is a *negative* example if $t(x) = 0$. The training sample may conveniently be regarded as an element of the product space $(X \times \{0, 1\})^m$,

$$\mathbf{x}(t) = ((x_1, b_1), (x_2, b_2), \dots, (x_m, b_m))$$

where, for each i , $b_i = t(x_i)$. The learning algorithm \mathcal{L} (which may be deterministic or randomised) produces an *output hypothesis* $\mathcal{L}(\mathbf{x}(t)) \in H$ which is to be taken as an approximation, in some sense, to the target. Ideally, we have exact learning of the target, in which case the output hypothesis is the same as the target. In general, if $h \in H$ and if $h(x_i) = t(x_i)$ for $1 \leq i \leq m$, then h is said to be *consistent* with t on the sample \mathbf{x} . Clearly, if \mathcal{L} always outputs a consistent hypothesis and if the only hypothesis consistent with a sample \mathbf{x} is t itself, then the learning algorithm exactly identifies t . If \mathbf{x} has this property, we say that \mathbf{x} *specifies* t (in H), and that it is a *specifying sample* for t . We also say that the set of examples in \mathbf{x} specifies t . In this case, no hypothesis in H distinct from the target agrees with the target on these examples. (Goldman and Kearns [11] have used the terminology *teaching sequence* for what we call a specifying sample, while Shinohara and Miyano [24] have used the term *key*.) The following parameter quantifies the difficulty of specifying a given hypothesis.

Definition 1.1 *Let H be a finite hypothesis space. Then for $t \in H$, the specification number of t (in H) is the least length of a specifying sample for t in H . \square*

The specification number of t in H is denoted by $\sigma_H(t)$, or simply by $\sigma(t)$ when H is clear.

1.2 Overview

We start by proving that the specification number of any hypothesis is at least equal to the *testing dimension* of Romanik and Smith [21].

We study in some depth the specification of the space H_n of linearly separable Boolean functions of n variables. These functions are those for which the positive examples can be separated from the negative examples by a hyperplane. We present two methods for obtaining upper bounds on the specification numbers. We show that hypotheses with small numbers of positive or negative examples are easily specified and we give a characterisation of those t with largest possible value of $\sigma(t)$ (that is, the most difficult hypotheses to teach). We give a tight upper bound on $\sigma(t)$ when t depends on all n co-ordinates (that is, there are no irrelevant variables). We prove that $\sigma(t) \geq n + 1$ for all t , and that equality holds if t is *nested*. We then give a simple proof that for any t there is a *unique* set of $\sigma(t)$ examples which specifies t (a fact noted by Cover [9]; see also [14]). Next, by a *projection* method, we determine an expression for $\sigma(t)$ when t depends only on a certain number of the n co-ordinates. For every $k \leq n$, we give a tight lower bound and a tight upper bound on $\sigma(t)$ for hypotheses depending on k co-ordinates. We then prove that, on average, only at most n^2 examples are needed to specify a hypothesis in H_n ; that is, the expected specification number is at most n^2 .

Finally, we discuss the complexity of computing specification numbers and teaching dimensions. We show that these problems are NP-hard, and remain so when restricted to *trivalent* hypothesis spaces.

1.3 Related Work

A recent paper by Goldman and Kearns [11] provides some calculations of specification numbers, for various types of hypotheses. In that paper, they define the *teaching dimension* of a hypothesis space H . In our terminology, this is the maximum, over all hypotheses of H , of the specification number, $\text{TD}(H) = \max\{\sigma(t) : t \in H\}$.

In [12], Goldman, Kearns and Schapire discuss a related idea: a *universal teaching sequence* for H is a sample \mathbf{x} such that \mathbf{x} specifies each hypothesis in H . That is to say, if $t, h \in H$ and t and h agree on \mathbf{x} then $t = h$.

Shinohara and Miyano [24] have studied what we call specification and have described algorithms for producing small specifying samples for the space of Boolean threshold functions (and for monomials). They have also studied the complexity of finding small specifying samples (or *keys*, as they call them). In their paper, they relate models of teaching and models of learning.

Salzberg *et al.* [23] have considered ‘learning with a helpful teacher’ when the learner uses a particular algorithm, which is known to the teacher. Specifically, they consider

a learner using the nearest-neighbour classification algorithm and a teacher trying to teach various types of geometric concepts. In our model, the teacher knows nothing of the algorithm the learner is applying; it is for this reason that we use the term *specification* rather than *teaching*.

Romanik and Smith [21, 22], in studying not exact specification, but approximate testing of geometric hypotheses, have introduced the *testing dimension*, $\tau(H)$, of a hypothesis space H . This is the maximum integer k such that *all* subsets of X of cardinality k are shattered by H ; that is, if $K \subseteq X$ has cardinality k , then all 2^k possible classifications of K into positive and negative examples can be achieved by hypotheses of H . (The testing dimension will in general be far less than the Vapnik-Chervonenkis dimension [26], a parameter which has proven to be of great importance in learning theory [5]. The VC dimension of H is defined as the maximum integer k such that *there is* a set of k examples shattered by H .) We have the following relation.

Proposition 1.2 *For any hypothesis space H and any $t \in H$, $\sigma_H(t) \geq \tau(H)$.*

Proof: Let \mathbf{x} be any sample of length less than $\tau(H)$, and let y be any example not contained in \mathbf{x} . Then, since H shatters the set consisting of y and the examples in \mathbf{x} , there is $h \in H$ such that h and t agree on \mathbf{x} , but such that $h(y) \neq t(y)$. Hence \mathbf{x} does not specify t , and the result follows. \square

2 Specifying Linearly Separable Functions

2.1 Introduction

In this section, we discuss the class of *linearly separable* Boolean functions on n variables. A Boolean function t defined on $\{0, 1\}^n$ is linearly separable if there are $\alpha \in \mathbf{R}^n$ and $\theta \in \mathbf{R}$ such that

$$t(x) = \begin{cases} 1 & \text{if } \langle \alpha, x \rangle \geq \theta \\ 0 & \text{if } \langle \alpha, x \rangle < \theta, \end{cases}$$

where $\langle \alpha, x \rangle$ is the standard inner product of α and x . Given such α and θ , we say that t is represented by $[\alpha, \theta]$ and we write $t \leftarrow [\alpha, \theta]$. The vector α is known as the *weight-vector*, and θ is known as the *threshold*. This class of functions is the set of functions computable by the *Boolean perceptron*, and we shall denote it by H_n . Of course, each $t \in H_n$ will satisfy $t \leftarrow [\alpha, \theta]$ for ranges of α and θ . A technical point, which will prove useful in some of the following analysis, is that, since the examples are discrete, for any $t \in H_n$, there are $\alpha \in \mathbf{R}^n$, $\theta \in \mathbf{R}$ and a positive constant c such that

$$t(x) = 1 \implies \langle \alpha, x \rangle \geq \theta + c, \quad t(x) = 0 \implies \langle \alpha, x \rangle \leq \theta - c.$$

Henceforth, for ease of notation, we denote the specification number of $t \in H_n$ in H_n by $\sigma_n(t)$.

2.2 Upper bounds

Clearly, the teaching dimension of H_n is at most 2^n , the total number of examples. In fact, it is easy to see that it is this bad.

Proposition 2.1 *The teaching dimension of H_n is 2^n .*

Proof: Observe that the identically-0 function ξ is in H_n . But so also are the hypotheses with precisely one positive example. If an example x is not presented, then the hypothesis which has just x as a positive example has not been ruled out. Hence to specify ξ , all 2^n examples must be presented, and $\sigma_n(\xi) = 2^n$. \square

We may be more specific and ask what the specification numbers of other, more interesting, linearly separable Boolean functions are. First, we present some general techniques for bounding these specification numbers.

For our first approach, we regard the points in $\{0,1\}^n$ as the vertices of the n -dimensional Boolean hypercube. For any $t \in H_n$, $\text{pos}(t)$, the set of positive examples of t and $\text{neg}(t)$, the set of negative examples of t , are connected subsets (in the graph-theoretic sense) of the hypercube. Let $\Delta(t)$ be the set of examples x such that there is y adjacent to x with $t(x) \neq t(y)$. Thus, $\Delta(t)$ may be thought of as the examples on the *boundary* between $\text{pos}(t)$ and $\text{neg}(t)$. Let us denote the cardinality of $\Delta(t)$ by $\partial(t)$. Then we have the following bound.

Proposition 2.2 (Boundary Result) *For $t \in H_n$, not the identically-0 hypothesis or the identically-1 hypothesis, $\sigma_n(t) \leq \partial(t)$, where $\partial(t)$ is the number of boundary examples of t .*

Proof: If t is non-constant, then it has boundary examples. Any sample which contains the examples in $\Delta(t)$ is a specifying sample for t , since such a sample certainly delineates the boundary between positive and negative examples. \square

The following consequence of this result is useful for hypotheses with small numbers of positive examples or negative examples.

Proposition 2.3 *For $t \in H_n$, not the identically-0 or the identically-1 hypothesis, let*

$$m = \min(|\text{pos}(t)|, |\text{neg}(t)|).$$

Then $\sigma_n(t) \leq m(n-1) + 2$.

Proof: Suppose, without loss, that $m = |\text{pos}(t)|$. Let P be the subgraph of the Boolean hypercube vertex-induced by $\text{pos}(t)$. Then P is connected, and so has at least

$m - 1$ edges. It follows, since each positive example has n neighbours, that the number of boundary vertices satisfies $\partial(t) \leq m + mn - 2(m - 1) = m(n - 1) + 2$. \square

We have seen that the identically-0 function ξ has specification number 2^n . Consider the hypotheses of the form $t(a_1, a_2, \dots, a_n) = 1 \iff a_i = b$, where i is an integer between 1 and n and b is 0 or 1. We call such hypotheses *hyperface hypotheses*. It is easily seen that any hyperface hypothesis has specification number 2^n . As an application of the Boundary Result, we can characterise the hypotheses of H_n which have largest possible specification numbers—that is, those which may be regarded as the most difficult to teach.

Proposition 2.4 *If $t \in H_n$ has $\sigma_n(t) = 2^n$ then t is either the identically-0 function, the identically-1 function, or a hyperface hypothesis.*

Proof: Suppose t is neither the identically-0 function nor the identically-1 function and that $\sigma_n(t) = 2^n$. Then, by Proposition 2.2, $\partial(t) = 2^n$, and all 2^n examples are boundary examples of $t \leftarrow [\alpha, \theta]$. Without loss, assume $\alpha_i \geq 0$ for $1 \leq i \leq n$ and that $\theta > 0$. (This is an assumption we can make without loss since it is not important which point we define to be the origin; thus, we may take as the origin the negative example furthest from the defining hyperplane.) The negative example $(00 \dots 00)$ is a boundary example of t , and so for some i the example $(00 \dots 010 \dots 0)$ with a 1 in the i th co-ordinate, is a positive example, whence $\alpha_i \geq \theta$. Also, since the positive example $(11 \dots 11)$ is a boundary example, for some j , the example $(11 \dots 101 \dots 1)$ with a 0 in the j th co-ordinate, is a negative example of t . Since $\alpha_i \geq \theta$, we must have $i = j$. Then $\alpha_i \geq \theta$ and $\sum_{j \neq i} \alpha_j < \theta$, so that t is the hyperface hypothesis $a_i = 1$. \square

For our second approach to bounding specification numbers, we regard the vertices of the hypercube as corresponding to all subsets of an n -set (so that the origin corresponds to the empty set and the examples with k entries equal to 1 correspond to the k -subsets.) Then the examples in $\{0, 1\}^n$ have a partial order \preceq on them, induced by inclusion in the power set lattice of the n -set. In this partial order on $\{0, 1\}^n$, $x \preceq y$ if $(x)_i = 1$ implies $(y)_i = 1$. This is quite different from the partial order used by Hu [14]. We say that t depends on co-ordinate i if there are $x^{(0)}, x^{(1)}$ differing only in their i th entries, such that $t(x^{(0)}) = 0, t(x^{(1)}) = 1$. In this case, the sign of α_i can be determined from $x^{(0)}$ and $x^{(1)}$ since $\langle \alpha, x^{(1)} \rangle \geq \theta > \langle \alpha, x^{(0)} \rangle$. Suppose that t depends on all the co-ordinates. Then we may, without loss, suppose that t is *increasing*, by which we mean that $t(x) = 1$ and $x \preceq y$ imply $t(y) = 1$. (We can assume without any loss that the hypothesis is increasing because the origin can be taken to be any point and we can determine from t which point is acting as the origin.) Then $\text{neg}(t)$ forms a down-set and $\text{pos}(t)$ an up-set with respect to \preceq . Let $D(t)$ be the set of maximal elements in $\text{neg}(t)$ and $U(t)$ the set of minimal elements in $\text{pos}(t)$.

Theorem 2.5 *Suppose $t \in H_n$ is increasing and depends on all the co-ordinates. Then the set $D(t) \cup U(t)$ specifies t .*

Proof: Suppose $t \leftarrow [\alpha, \theta]$. We claim that the signs of $\alpha_1, \alpha_2, \dots, \alpha_n$ can be deduced solely from the classification of $D(t) \cup U(t)$ and the fact that $t \in H_n$. Suppose without loss of generality that $0 < \alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n$. We prove, by induction on m , that the signs of $\alpha_1, \alpha_2, \dots, \alpha_m$ can be deduced solely from the classification of the examples in $D(t) \cup U(t)$ and the knowledge that $t \in H_n$. The statement is trivial for $m = 0$, so we move on to the induction step. Suppose that $\alpha_1, \dots, \alpha_{m-1}$ are given to be positive; we show that it can be deduced that α_m is positive. Since t depends on co-ordinate m , there are y, y' , differing only in that $(y)_m = 0$ and $(y')_m = 1$, such that $\langle \alpha, y \rangle < \theta$ and $\langle \alpha, y' \rangle \geq \theta$. Then $y \preceq x$ for some $x \in D(t)$ and $(x)_m = 0$ since otherwise $y' \preceq x$. Let x' be the example equal to x except that $(x')_m = 1$; then x' is a positive example, since $x \preceq x'$ and $x \in D(t)$. Take $z' \in U(t)$ with $z' \preceq x'$ and let z be equal to z' except that $(z)_m = 0$. Then,

$$\langle \alpha, z' \rangle \geq \theta > \langle \alpha, x \rangle = \langle \alpha, x' \rangle - \alpha_m.$$

So $\langle \alpha, x' - z' \rangle < \alpha_m$, and hence x' and z' do not differ in co-ordinates $m+1, m+2, \dots, n$. Let C be the set of co-ordinates i such that $(x')_i \neq (z')_i$. Now we have

$$0 < \langle \alpha, z' \rangle - \langle \alpha, x \rangle = \alpha_m - \sum_{i \in C} \alpha_i.$$

The above inequality can be deduced solely from the facts that $z' \in U(t)$ and $x \in D(t)$. Thus, given additionally the information that $\alpha_i > 0$ for $i \in C \subseteq \{1, \dots, m-1\}$, it can be deduced that $\alpha_m > 0$ also. Therefore, given only the classification of $D(t) \cup U(t)$, one can deduce that t is increasing. Furthermore, t is specified by $D(t) \cup U(t)$: for any example y , there will be a point x in $D(t)$ with $y \preceq x$ or there will be $z \in U(t)$ with $z \preceq y$. In the first case, $t(y) = 0$ and in the second $t(y) = 1$. \square

As an immediate corollary of this result, we have the following bound.

Corollary 2.6 *If $t \in H_n$ depends on all the co-ordinates then*

$$\sigma_n(t) \leq \binom{n+1}{\lfloor \frac{n+1}{2} \rfloor}.$$

Proof: We may, without loss, suppose that t is increasing. (Clearly if t depends on all the co-ordinates but is not increasing, one may simply shift the origin to yield an analogous specifying set. The teacher knows where the origin should be shifted to, so can effect this transformation. Equivalently, the teacher transforms the order \preceq and produces as a specifying sample the minimal positive examples and maximal negative examples with respect to the transformed ordering.) Then $\sigma_n(t) \leq |D(t)| + |U(t)|$. Now form a set A consisting of all points $x1$ for $x \in D(t)$ and all points $y0$, for $y \in U(t)$. We now show that A is an antichain in the poset $(\{0, 1\}^{n+1}, \preceq)$. It is clear that, since the elements of $D(t)$ are incomparable, so are the elements of the form $x1$ where $x \in D(t)$. Similarly, the points $y0$ for $y \in U(t)$ are incomparable. Also, for any $x \in D(t)$ and $y \in U(t)$, $x1$ and $y0$ are incomparable since it cannot be true that $y \preceq x$ (since t is increasing). Sperner's Theorem [25, 6, 1] shows that the maximal size of an antichain in $(\{0, 1\}^{n+1}, \preceq)$ is the quantity stated, and the result follows since $|A| = |D(t) \cup U(t)|$. \square

Let g_n^k be the hypothesis which has as positive examples the examples with at least k ones. Then $g_n^k \leftarrow [(1, 1, \dots, 1), k] \in H_n$, and we shall call it the *weight-at-least- k* hypothesis.

Proposition 2.7 *The weight-at-least- k hypothesis g_n^k has specification number $\binom{n+1}{k}$.*

Proof: If x, y are adjacent vertices of the hypercube then their weights (number of ones) differ by 1. Hence x, y are adjacent and $g_n^k(x) = 1, g_n^k(y) = 0$ if and only if x has weight k and y has weight $k - 1$. It follows that the set $\Delta(g_n^k)$ consists of all examples with weight k and all examples of weight $k - 1$, so that

$$\sigma_n(g_n^k) \leq \partial(g_n^k) = \binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}.$$

We now show all examples of weight k and of weight $k - 1$ must be presented to specify g_n^k . (Then every specifying sample must contain all such examples, yielding the lower bound.) Let x be an example of weight $k - 1$, which we may suppose is $x = (11 \dots 10 \dots 0)$. Let $\alpha = (11, \dots, 11)$, and let $\theta = k$. Then $g_n^k \leftarrow [\alpha, \theta]$. Now let

$$\beta = \underbrace{(1 + 1/k, 1 + 1/k, \dots, 1 + 1/k)}_{k-1}, 1, 1, \dots, 1$$

and note that $g_n^k \leftarrow [\beta, k]$. Now, $h \leftarrow [\beta, k - 1/k]$ misclassifies x as a positive example and correctly classifies all other examples. Hence x must be presented if g_n^k is to be specified. The treatment for examples of weight k is similar. \square

This shows that we can have equality in Corollary 2.6, achieved by the weight-at-least- $\lfloor (n+1)/2 \rfloor$ hypothesis. In fact, we can characterise precisely those linearly separable functions which depend on all the variables and have highest specification number.

Proposition 2.8 *Suppose $t \in H_n$ depends on all the co-ordinates. Then t has maximum possible specification number for such a hypothesis if and only if one of the following holds:*

- (i) n is odd and there is $v \in \{0, 1\}^n$ such that $t(x) = 1$ if and only if x and v agree on at least $(n+1)/2$ entries;
- (ii) n is even and there is $v \in \{0, 1\}^n$ such that $t(x) = 1$ if and only if x and v agree on at least $n/2$ entries;
- (iii) n is even and there is $v \in \{0, 1\}^n$ such that $t(x) = 1$ if and only if x and v agree on at least $(n/2 + 1)$ entries.

Proof: Consider the lattice of subsets of an n -set. It is known (see [1, 6]) that if n is even then there is exactly one antichain of the maximum possible size $\binom{n}{n/2}$ — namely,

the collection of all subsets of cardinality $n/2$. For n odd, the only antichains of size $\binom{n}{\lfloor n/2 \rfloor}$ are the collection of all $(n-1)/2$ -subsets and the collection of all $(n+1)/2$ -subsets. Consider the proof of Corollary 2.6 in the case n odd. Then we have equality in the bound if and only if the antichain A is the set of all examples of weight $(n+1)/2$, from which it follows that the maximal negative examples of t are all the examples of weight $(n-1)/2$ and the minimal positive examples are all the examples of weight $(n+1)/2$. But this means that if n is odd, if t is increasing and depends on all the co-ordinates and we have equality in Corollary 2.6 then t must be the weight-at-least- $(n+1)/2$ hypothesis. Conversely, such a hypothesis meets the upper bound. If t is not increasing then we may take some other point y as the origin to transform t to an increasing function. Then t is as described in the statement of this result, with $v = (1, 1, \dots, 1) - y$. The proof for n even is similar. \square

Theorem 2.5 shows that for increasing t depending on all the co-ordinates, the specification number is at most $|D(t) \cup U(t)|$. It is worth remarking that, in general, one does not have equality here. That is, the specification number can be strictly smaller. It is easy to see also that the specification number can easily be less than the number of boundary examples.

2.3 A lower bound and its attainment

We have characterised the hypotheses in H_n with largest specification numbers. Now we turn our attention to those hypotheses with the lowest possible specification numbers. The testing dimension of H_n is just 3, so we cannot obtain any useful lower bound using this approach. However, a straightforward lower bound can easily be obtained. We say that a set of $n+1$ points in \mathbf{R}^n is in *general position* if the points do not all lie on a hyperplane.

Theorem 2.9 *For any $t \in H_n$, any specifying sample for t contains $n+1$ examples in general position, and possibly some others. In particular, $\sigma_n(t) \geq n+1$. Furthermore, equality can hold in this bound.*

Proof: Suppose that T is a set of examples not containing $(n+1)$ points in general position. Then all the points of T lie in some hyperplane with equation $\langle \beta, x \rangle = c$, for some $\beta \in \mathbf{R}^n$ and $c \in \mathbf{R}$. Let $t \leftarrow [\alpha, \theta]$ be any hypothesis in H_n and let η be any real number. Then if $h_\eta \leftarrow [\alpha + \eta\beta, \theta + \eta c]$, h_η agrees with t on T ; for, if $x \in T$ then $\langle \beta, x \rangle = c$ and

$$\langle \alpha + \eta\beta, x \rangle = \langle \alpha, x \rangle + \eta\langle \beta, x \rangle = \langle \alpha, x \rangle + \eta c,$$

whence, for $x \in T$,

$$t(x) = 1 \iff \langle \alpha, x \rangle \geq \theta \iff \langle \alpha + \eta\beta, x \rangle \geq \theta + \eta c.$$

Now, choose $y \in \{0, 1\}^n$ which does not lie on the flat determined by T , so that $\langle \beta, y \rangle \neq c$. Then for some values of η , y is a negative example of h_η and for some other

values, y is a positive example of h_η . In other words, the sample T does not specify t , since there are at least two distinct hypotheses consistent with t on the sample. To see that the lower bound can be attained, note that if t has just one positive example, then the sample consisting of this positive example and its n neighbours is a sample of length $n + 1$ which specifies t . \square

The fact that at least $n + 1$ examples are required has been shown in [14], but the proof presented here is more direct. We have seen that the hypotheses having exactly one positive example or one negative example have the least possible specification number $n + 1$. But these are not the only such hypotheses, as we now show.

Using the standard formula notation in terms of the literals u_1, u_2, \dots, u_n (and their negations), let us recursively define a Boolean function to be *nested* by: both functions of 1 variable are nested, and t_n , a function of n variables, is nested if $t_n = u_n \star t_{n-1}$ or $t_n = \bar{u}_n \star t_{n-1}$ where \star is \vee (the OR connective) or \wedge (the AND connective) and t_{n-1} is a nested function of $n - 1$ variables. (Here, we mean that t_{n-1} acts on the first $n - 1$ entries of its argument.) Any nested Boolean function is linearly separable. For, if $t_{n-1} \leftarrow [\alpha, \theta]$ is nested and in H_{n-1} , then

$$\begin{aligned} u_n \wedge t_{n-1} &\leftarrow [(\alpha, M), \theta + M], & u_n \vee t_{n-1} &\leftarrow [(\alpha, M), \theta], \\ \bar{u}_n \wedge t_{n-1} &\leftarrow [(\alpha, -M), \theta], & \bar{u}_n \vee t_{n-1} &\leftarrow [(\alpha, -M), \theta - M], \end{aligned}$$

for a suitably large M . Examples of nested hypotheses include the hypotheses with formulae $u_1 \wedge u_2 \wedge \dots \wedge u_n$ and $u_1 \vee u_2 \vee \dots \vee u_n$, with only one positive example, and one negative example (respectively), and the hypothesis

$$f_n = u_n \wedge (u_{n-1} \vee (u_{n-2} \wedge (u_{n-3} \vee (\dots u_1) \dots))),$$

which is of interest in the context of the perceptron learning algorithm [13, 18, 4]. The next result shows that all nested hypotheses are easily specified.

Theorem 2.10 *The specification number of any nested hypothesis in H_n is $n + 1$.*

Proof: It suffices to prove that for any increasing nested hypothesis in H_n , $|D(t) \cup U(t)| \leq n + 1$. This is clearly true when $n = 1$, for in this case the total number of examples is 2. Suppose the statement is true for $(n - 1) \geq 1$, and consider n . Let t_n be nested in H_n . If $t_n = u_n \vee t_{n-1}$, then $D(t_n)$ consists of all examples $x0$ where $x \in D(t_{n-1})$, and $U(t_n)$ consists of all examples $y0$ where $y \in U(t_{n-1})$ together with the single example $(00 \dots 01)$. If $t_n = u_n \wedge t_{n-1}$ then $D(t_n)$ consists of all examples $x1$ where $x \in D(t_{n-1})$, together with the example $(11 \dots 10)$, and $U(t_n)$ consists of the examples $y1$ where $y \in U(t_{n-1})$. In either case, $\sigma_n(t_n) \leq |D(t_n)| + |U(t_n)| \leq 1 + |D(t_{n-1})| + |U(t_{n-1})| = 1 + n$, as required. \square

We may extend the definition of nested hypothesis by allowing the variables to be permuted (or re-labelled), so that we would say for example that the function $u_2 \wedge$

$(u_3 \vee u_1)$ is nested. Clearly the above result is true for this more general definition of a nested hypothesis.

One may relate nested hypotheses to particular types of *decision lists*, introduced by Rivest [20]. It is straightforward to show (inductively) that any nested hypothesis can be realised as a 1-decision list of length n . Conversely, with the more general definition of nested hypothesis in which we allow the variables to be re-labelled, any 1-decision list of length n computes a nested hypothesis.

We conjecture that the only hypotheses in H_n which have specification number $n + 1$ are the nested hypotheses.

2.4 Signatures

In calculating the specification number of the weight-at-least- k hypothesis, we used the fact that if x has the property that there is $h \in H_n$ with $h(y) = t(y)$ for all $y \neq x$ and $h(x) \neq t(x)$ then x must belong to any specifying sample for t . We shall say that an example with this property is *essential* for t . (Cover [9] describes such examples as *ambiguous*.) Clearly any specifying sample for t must contain all examples which are essential for t . We now give a simple proof of the fact, first observed by Cover [9], that the essential examples alone are sufficient to specify. (Cover did not present a proof of the result in the paper cited, but refers to work of Mays [19] on boundary matrices. Hu [14] presents a proof based on the work of Mays [19].)

Theorem 2.11 *Let $t \in H_n$ and let $S(t)$ be the set of examples essential for t . Then $S(t)$ specifies t .*

Proof: Suppose not. Then there is h agreeing with t on $S(t)$ but disagreeing on some other examples. Let's say $t \leftarrow [\alpha, \theta]$ and $h \leftarrow [\beta, \phi]$, where no example lies on the defining hyperplanes. For $0 \leq \lambda \leq 1$, consider the hypothesis

$$\lambda t + (1 - \lambda)h \leftarrow [\lambda\alpha + (1 - \lambda)\beta, \lambda\theta + (1 - \lambda)\phi].$$

The hypothesis $\lambda t + (1 - \lambda)h$ correctly classifies any example in $S(t)$ since each of h, t correctly classifies such examples. Suppose y is misclassified by h . Then $\langle \alpha, y \rangle > \theta$ and $\langle \beta, y \rangle < \phi$, or $\langle \alpha, y \rangle < \theta$ and $\langle \beta, y \rangle > \phi$. The function

$$f(\lambda) = \lambda\langle \alpha, y \rangle + (1 - \lambda)\langle \beta, y \rangle - \lambda\theta - (1 - \lambda)\phi$$

is continuous and strictly increasing or strictly decreasing and $f(0), f(1)$ are of opposite signs, so there is a unique $0 < \lambda_y < 1$ such that

$$\lambda_y\langle \alpha, y \rangle + (1 - \lambda_y)\langle \beta, y \rangle = \lambda_y\theta + (1 - \lambda_y)\phi.$$

Furthermore, it is easy to see that α, β could have been chosen in such a way that if $y \neq z$ then $\lambda_y \neq \lambda_z$. Observe that if $\lambda_y > \lambda_z$ then the hypothesis $\lambda_y t + (1 - \lambda_y)h$

correctly classifies z . Now, since the λ_x are distinct, there is some example v such that $\lambda_v > \lambda_y$ for all $y \neq v$ misclassified by h . Thus (by taking a value of λ very close to λ_v), there is a hypothesis $\lambda t + (1 - \lambda)h$ which classifies all examples but v correctly. Therefore $v \in S(t)$. But this is a contradiction, since we assumed h to be consistent with t on $S(t)$ (in which case, any such convex combination of t and h would classify v correctly). \square

Corollary 2.12 *Let $t \in H_n$. Then there is precisely one set of $\sigma_n(t)$ examples which specifies t . This set is $S(t)$.* \square

We shall call the set $S(t)$ of all examples essential for t the *signature* of t . Any specifying sample contains these examples, and so the signature is the unique minimal specifying set for t .

2.5 Projections

From any $t \in H_n$, we can form two hypotheses $t \uparrow, t \downarrow$ of H_{n-1} , as follows:

$$t \uparrow (a_1, a_2, \dots, a_{n-1}) = t(a_1, a_2, \dots, a_{n-1}, 1),$$

$$t \downarrow (a_1, a_2, \dots, a_{n-1}) = t(a_1, a_2, \dots, a_{n-1}, 0).$$

Thus, $t \uparrow$ is the restriction of t to the hyperface $x_n = 1$ of the Boolean hypercube and $t \downarrow$ is its restriction to the hyperface $x_n = 0$. We call $t \uparrow$ the *up-projection* and $t \downarrow$ the *down-projection* of t . Note that if $t \leftarrow [\alpha, \theta]$ where $\alpha = (\beta, d)$, $\beta \in \mathbf{R}^{n-1}$ and $d \in \mathbf{R}$, then $t \uparrow \leftarrow [\beta, \theta - d]$ and $t \downarrow \leftarrow [\beta, \theta]$.

Theorem 2.13 (Projection Result) *For $t \in H_n$,*

$$\sigma_n(t) \leq \sigma_{n-1}(t \uparrow) + \sigma_{n-1}(t \downarrow),$$

and equality holds when $t \uparrow = t \downarrow$.

Proof: It is easy to see that the inequality holds. Let $S(t \downarrow)$ be the signature of $t \downarrow$ and $S(t \uparrow)$ the signature of $t \uparrow$. For each $s = (a_1, a_2, \dots, a_{n-1})$ in $S(t \downarrow)$, form the example $s0 = (a_1, a_2, \dots, a_{n-1}, 0)$ and for each $s = (a_1, a_2, \dots, a_{n-1}) \in S(t \uparrow)$, form the example $s1 = (a_1, a_2, \dots, a_{n-1}, 1)$. Then it is clear that these examples specify t , so that

$$\sigma_n(t) \leq |S(t \downarrow)| + |S(t \uparrow)| = \sigma_{n-1}(t \downarrow) + \sigma_{n-1}(t \uparrow).$$

In order to prove equality when $t \uparrow = t \downarrow$, we first prove that if z is any point in the signature $S(t)$ of t and $t(z) = 1$ (resp., $t(z) = 0$) then there are α, θ such that $t \leftarrow [\alpha, \theta]$ and for any other positive (resp., negative) example x of t , $\langle \alpha, x \rangle > \langle \alpha, z \rangle$ (resp.,

$\langle \alpha, x \rangle < \langle \alpha, z \rangle$). Suppose $z \in S(t)$ is a positive example of t . Then there is $h \leftarrow [\beta, \phi]$ such that h agrees with t on all examples except z , and such that no example x satisfies $\langle \beta, x \rangle = \phi$. We may assume (by the comment at the end of Section 2.1) that there is $c > 0$ such that $\langle \alpha, z \rangle \geq \theta + c$, $\langle \beta, z \rangle \leq \phi - c$ and such that for $x \in \text{neg}(t)$, $\langle \alpha, x \rangle \leq \theta - c$ and $\langle \beta, x \rangle \leq \phi - c$, and for $z \neq x \in \text{pos}(t)$, $\langle \alpha, x \rangle \geq \theta + c$ and $\langle \beta, x \rangle \geq \phi + c$. Let λ be such that

$$\lambda \langle \alpha, z \rangle + (1 - \lambda) \langle \beta, z \rangle = \lambda \theta + (1 - \lambda) \phi.$$

Let $\gamma = \lambda \alpha + (1 - \lambda) \beta$ and $\psi = \lambda \theta + (1 - \lambda) \phi$. Then, as can easily be checked, $[\gamma, \psi]$ represents t . Further, for $z \neq x \in \text{pos}(t)$,

$$\langle \gamma, x \rangle = \langle \lambda \alpha + (1 - \lambda) \beta, x \rangle \geq \lambda \theta + (1 - \lambda) \phi + c > \lambda \theta + (1 - \lambda) \phi = \langle \gamma, z \rangle.$$

The argument when z is a negative example is similar.

Now we show that the set S consisting of all examples $z1$ and $z0$, for $z \in S(t \downarrow)$ is the signature of t . As mentioned above, the points of S specify t . We prove that all are essential for t , from which it follows that $S = S(t)$. Without loss of generality, suppose $s = z1$ where z is a positive example of $t \downarrow$. We wish to find $h \leftarrow [\beta, \phi]$ where $\beta \in \mathbf{R}^n, \phi \in \mathbf{R}$ such that $h(z0) = 1, h(z1) = 0, h(x1) = h(x0) = 1$ for all $z \neq x \in \text{pos}(t \downarrow)$, and $h(x1) = h(x0) = 0$ for all $x \in \text{neg}(t \downarrow)$. By the above, we may assume that $t \downarrow \leftarrow [\alpha, \theta]$, where for some $c > 0$,

$$z \neq x \in \text{pos}(t \downarrow) \implies \langle \alpha, x \rangle \geq \langle \alpha, z \rangle + c.$$

Let $\beta = (\alpha, -c)$ and let $\phi = \langle \alpha, z \rangle$. Then $\langle \beta, z1 \rangle = \langle \alpha, z \rangle - c < \phi$ and $\langle \beta, z0 \rangle = \langle \alpha, z \rangle$, so that $h(z1) = 0$ and $h(z0) = 1$. For $x \in \text{pos}(t \downarrow), x \neq z$, we have

$$\langle \beta, x0 \rangle \geq \langle \beta, x1 \rangle = \langle \alpha, x \rangle - c \geq \langle \alpha, z \rangle = \phi,$$

whence $h(x0) = h(x1) = 1$. Also, for $x \in \text{neg}(t \downarrow)$,

$$\langle \beta, x1 \rangle \leq \langle \beta, x0 \rangle = \langle \alpha, x \rangle < \theta,$$

which is less than ϕ since $\langle \alpha, z \rangle > \theta$ and so $h(x1) = h(x0) = 0$. The result follows. \square

We now turn our attention to hypotheses in H_n which depend on a particular number, k , of the co-ordinates. Such a hypothesis has $n - k$ ‘irrelevant attributes’ (as defined in [18]). Suppose that t depends on co-ordinates 1 to k only and denote by t_k the function in H_k defined by $t_k((a_1, a_2, \dots, a_k)) = t((a_1, a_2, \dots, a_k, 0, 0, \dots, 0))$, obtained by projecting t down $n - k$ times. Then we have the following result, an immediate consequence of the Projection Result.

Proposition 2.14 *If $t \in H_n$ and t depends only on co-ordinates 1, 2, \dots , k , then the specification number of t equals $2^{n-k} \sigma_k(t_k)$. \square*

As an example of this, consider the hypothesis $g \in H_n$ defined by $g(x) = 1$ if and only if, of the first k entries of x , at least r are equal to 1. Then g is the r -out-of- k hypothesis and is easily seen to be linearly separable. Clearly, g depends only

on the first k co-ordinates and $g_k \in H_k$ is the weight-at-least- r hypothesis, so that $\sigma_n(g) = 2^{n-k} \sigma_k(g_k) = 2^{n-k} \binom{k+1}{r}$.

We have the following tight bound, from Corollary 2.6, Theorem 2.8, and the Projection Result.

Theorem 2.15 *Suppose $t \in H_n$ depends on exactly k co-ordinates. Then*

$$2^{n-k}(k+1) \leq \sigma_n(t) \leq 2^{n-k} \binom{k+1}{\lfloor \frac{k+1}{2} \rfloor},$$

and equality is possible in both cases. □

From Proposition 2.8, it is easy to obtain a characterisation of those t meeting the upper bound above. Furthermore, our work on nested functions enables us to generate a class of hypotheses meeting the lower bound. Note that, since (as we mentioned earlier) any 1-decision list of length n ([20]) is a nested linearly separable Boolean function, any hypothesis realisable as a 1-decision list of length k has specification number $2^{n-k}(k+1)$ in the space H_n .

A consequence of Theorem 2.15 is that if a linearly separable Boolean function has few relevant attributes then the number of examples needed to specify it is exponential in n .

2.6 The expected specification number

We have now seen the extreme values that specification numbers in H_n can take. A natural problem is to determine the *average* or *expected* specification number, by which we mean the quantity

$$\overline{\sigma_n(t)} = \frac{1}{|H_n|} \sum_{t \in H_n} \sigma_n(t).$$

A set of N points in \mathbf{R}^n is said to be in *general position* if no $n+1$ of the points lie on a hyperplane. Given any such set X of points, we may define a set of $\{0, 1\}$ -valued functions on X by the same method we used to define the class of linearly separable Boolean functions; that is, for each hyperplane in \mathbf{R}^n , assign 1 to the points of X on and on one side of this hyperplane, and 0 to the others. Cover [9] has investigated such sets of functions. He proves that, asymptotically, the expected number of examples needed to specify one of these functions is $2(n+1)$. But Cover's analysis cannot be carried over to H_n , for here the set X is $\{0, 1\}^n$, a set of points certainly not in general position. (Indeed, it is easy to see that no set of $2n+1$ examples is in general position, for either $n+1$ of these lie on the hyperplane $x_1 = 0$ or $n+1$ lie on the hyperplane $x_1 = n$.) Therefore we must take an approach very different from that of Cover [9].

For the purposes of this section we will adapt the previous notation by incorporating the threshold as a weight. Hence a function $t \leftarrow [\alpha, \theta]$ will be represented by the *extended* weight vector $\alpha' = (\theta, \alpha_1, \dots, \alpha_n)$, while the examples will be augmented by a coordinate with value -1 . Hence example $x \in \{0, 1\}^n$ is represented by $x' = (-1, x_1, \dots, x_n)$. In this way we can write

$$t(x) = t(x') = \text{He}(\langle x', \alpha' \rangle),$$

where $\text{He}(z)$ is the *Heaviside function* given by

$$\text{He}(z) = \begin{cases} 1; & \text{if } z \geq 0; \\ 0; & \text{otherwise.} \end{cases}$$

Let $X \subseteq \{-1\} \times \{0, 1\}^n$ and consider a set of points

$$Y = \{y_1, \dots, y_k\} \subseteq (\{-1\} \times \{0, 1\}^n) \setminus X.$$

Let $H = H(X, Y)$ be the set of functions f on X such that there exist linearly separable functions f_1, \dots, f_{2k} which shatter y_1, \dots, y_k with the restriction $f_{i|X}$ of f_i to X equal to f for all i . (Recall that to say a set of functions F shatters a set of examples Y means that all possible classifications of Y into positive and negative examples can be realised by functions in F .) We say that $H(X, Y)$ is the set of linearly separable functions *restricted to X while shattering Y* . Note that if the examples in Y are not in general position then $H(X, Y) = \emptyset$, since they cannot be shattered at all. For the case when Y is in general position, if $|Y| > n + 1 = \text{VCdim}(H_n)$, the Vapnik-Chervonenkis dimension of H_n , then $|H(X, Y)| = 0$ since Y cannot be shattered, while if $|Y| = n + 1 = \text{VCdim}(H_n)$ then $|H(X, Y)| \leq 1$. To see this consider two distinct functions $f_1, f_2 \in H(X, Y)$ and choose an example $x \in X$ for which $f_1(x) \neq f_2(x)$. The extensions of f_1, f_2 together form a shattering set for $Y \cup \{x\}$, a contradiction.

Lemma 2.16 *If for some $x \in X$ and for some real numbers λ_y , $x = \sum_{y \in Y} \lambda_y y$, then $H(X, Y) = \emptyset$.*

Proof: Let $f \in H(X, Y)$ and consider the two extensions of f to linearly separable functions f_1, f_{-1} , such that for $y \in Y$, $f_1(y) = 1$ if and only if $\lambda_y > 0$, and $f_{-1}(y) = 1$ if and only if $\lambda_y < 0$. Suppose that f_1, f_{-1} are represented by (extended) weight vectors w_1, w_{-1} (respectively). Then

$$\langle w_1, x \rangle = \sum_{y \in Y} \lambda_y \langle w_1, y \rangle$$

and

$$\langle w_{-1}, x \rangle = \sum_{y \in Y} \lambda_y \langle w_{-1}, y \rangle$$

have different signs, so that $f_1(x) \neq f_{-1}(x)$. This is a contradiction, since both these functions are extensions of f . We conclude that $H(X, Y) = \emptyset$. \square

Consider the relation \sim on X defined as follows: $x_1 \sim x_2$ if and only if there are real numbers μ and λ_y , for each $y \in Y$, such that

$$x_1 = \mu x_2 + \sum_{y \in Y} \lambda_y y.$$

Lemma 2.17 *If $H(X, Y) \neq \emptyset$, the relation \sim is an equivalence relation.*

Proof: Since $H(X, Y) \neq \emptyset$, we conclude from the previous lemma that for any $x_1, x_2 \in X$, with $x_1 \sim x_2$, we have $\mu \neq 0$ in the equation

$$x_1 = \mu x_2 + \sum_{y \in Y} \lambda_y y.$$

Hence the relation is symmetric. It is also clearly reflexive: for $x_1 \sim x_2$ and $x_2 \sim x_3$, we combine the two equations to eliminate x_2 and obtain $x_1 \sim x_3$. \square

For sets X, Y , let X/Y be a set of representatives of the equivalence classes of X under the relation \sim . Note that if $|Y| = n$ and $H(X, Y) \neq \emptyset$, we have only one equivalence class since $Y \cup \{x\}$ forms a basis of \mathbf{R}^{n+1} for any $x \in X$ by Lemma 2.16.

Lemma 2.18 *Suppose $H(X, Y) \neq \emptyset$. There exists w such that $\langle w, y \rangle = 0$ for all $y \in Y$, $\langle w, x \rangle \neq 0$ for $x \in X$, and*

$$f(x) = \text{He}(\langle w, x \rangle),$$

for all $x \in X$, if and only if f belongs to $H(X, Y)$.

Proof: (\implies) Suppose we are given w satisfying the above conditions. Let Y^+ be any subset of Y . Choose δ_w such that

$$\begin{aligned} \langle \delta_w, y \rangle &= 1; \text{ for } y \in Y^+ \\ \text{and } \langle \delta_w, y \rangle &= -1; \text{ for } y \in Y \setminus Y^+. \end{aligned}$$

This can be done since this represents at most $n + 1$ linearly independent linear equations in $n + 1$ unknowns. Now consider

$$\hat{w}(\lambda) = w + \lambda \delta_w,$$

and $x \in X$. Since $\langle w, x \rangle \neq 0$, there exists $\epsilon_x > 0$ such that $\langle \hat{w}(\lambda), x \rangle \neq 0$ for $|\lambda| \leq \epsilon_x$. Let $\epsilon = \min_{x \in X}(\epsilon_x) > 0$. Taking $\hat{w} = \hat{w}(\epsilon)$ determines a linearly separable function which agrees with w on X and which computes 1 on Y^+ and 0 on $Y \setminus Y^+$. Since Y^+ was arbitrary, the function defined by w on X is in $H(X, Y)$.

(\impliedby) Suppose $f \in H(X, Y)$. Note first that any linearly separable function on a finite set X can be realised with a weight vector w such that

$$\langle w, x \rangle \neq 0, \text{ for } x \in X,$$

by slightly reducing the threshold if equality holds for some positively classified inputs. We prove the result by induction on $|Y|$. For $|Y| = 0$, by the above, there is nothing to prove. Assume now that the result holds in the case $|Y| = k - 1$ and let $Y = \{y_1, \dots, y_k\}$. Let f_i have weight vector w_i and assume that for $i \leq 2^{k-1}$, $f_i(y_k) = 0$, while for $i > 2^{k-1}$, $f_i(y_k) = 1$. Consider applying induction to the set $H(X \cup \{y_k\}, Y \setminus \{y_k\})$. We have two functions f^0, f^1 in this set agreeing with f on X and such that $f^0(y_k) = 0$ and $f^1(y_k) = 1$. By induction we can find w_0 for f^0 such that $\langle w_0, y_i \rangle = 0$ for $i = 1, \dots, k-1$, and for $x \in X \cup \{y_k\}$, $\langle w_0, x \rangle \neq 0$ with $\text{He}(\langle w_0, x \rangle) = f(x)$. So $\langle w_0, y_k \rangle < 0$. Likewise we find w_1 for f^1 . Taking $w(t) = tw_0 + (1-t)w_1$ we can choose t such that $\langle w(t), y_k \rangle = 0$ with $\langle w(t), x \rangle \neq 0$ and $\text{He}(\langle w(t), x \rangle) = f(x)$. \square

In view of this lemma, we introduce the following notation. For a weight vector w satisfying the conditions of the lemma, we denote the corresponding function in $H(X, Y)$ by f_w , while a weight vector obtained from a function $f \in H(X, Y)$ is denoted by w_f .

Lemma 2.19 *Consider sets $X, Y \subseteq \{-1\} \times \{0, 1\}^n$ and functions $H(X, Y)$ as above. Any specifying sample for $t \in H(X, Y)$ in $H(X, Y)$ can be replaced by a sample of the same length using only examples in X/Y . Hence*

$$\sigma_{H(X, Y)}(t) = \sigma_{H(X/Y, Y)}(t).$$

Proof: We simply replace any example x in the sample which is not in X/Y by the representative of its equivalence class. It will be sufficient to show that the value of any function in $H(X, Y)$ on x determines its value on x' when $x \sim x'$. This will also imply that the minimum specifying samples have the same length. Let

$$x = \mu x' + \sum_{y \in Y} \lambda_y y.$$

Consider any function $f \in H(X, Y)$ and let w_f be a weight vector guaranteed by Lemma 2.18. Since $\langle w_f, y \rangle = 0$ for all $y \in Y$, we have $\langle w_f, x \rangle = \mu \langle w_f, x' \rangle$. Hence if $\mu > 0$, $f(x) = f(x')$, while if $\mu < 0$, $f(x) \neq f(x')$. In either case the value of f on one of the two examples determines its value on the other, as required. \square

For a set of functions H and $t \in H$, the error set of a function $h \in H$ (with respect to t) is the set of examples on which t and h disagree. For a fixed target $t \in H$, we order the functions of H according to inclusion of their error sets. We will refer to the least non-empty sets in the inclusion ordering of the error sets as minimal error sets and to the corresponding functions as minimal error functions. In order to specify a target function t we must give a set of examples such that each minimal error set contains at least one of the examples. A special case occurs when the minimal error sets are singletons. In this case, as earlier, we call the examples in these sets *essential* and the list of essential examples is called the *signature* of t in H (denoted $S_H(t)$). In this case the signature is clearly the unique minimal specifying sample for t in H .

Using the machinery developed above, we are now ready to tackle our main task of describing specifying samples for linearly separable functions.

Proposition 2.20 *For any fixed $t \in H$, the corresponding minimal error functions in $H(X/Y, Y)$ have singleton error sets.*

Proof: We may assume that $|Y| < n$, since for $|Y| > n$, we have $|H(X, Y)| \leq 1$, while for $|Y| = n$, $|X/Y| = 1$. Suppose that t is the target and f is a minimal error function with error set containing x_1, x_2, \dots, x_m , $m > 1$. Let w_f be the weight vector guaranteed by Lemma 2.18 for f , and let w_t be the corresponding weight vector for t . Consider

$$w(\lambda) = (1 - \lambda)w_t + \lambda w_f.$$

For all examples not in the error set, the function $f_w(\lambda)$ will agree with both t and f . Since each x_i is differently classified by t and f , there exists λ_i such that $\langle w(\lambda_i), x_i \rangle = 0$. Let

$$\lambda_{\min} = \min\{\lambda_i : 1 \leq i \leq m\}.$$

Suppose $\lambda_i > \lambda_{\min}$. Taking

$$\lambda = (\min\{\lambda_j | \lambda_j \neq \lambda_{\min}\} + \lambda_{\min})/2,$$

the function $f_{w(\lambda)}$ lies strictly between t and f in the error sets ordering, contradicting the minimality of f . Hence $\lambda_i = \lambda_{\min}$ for all i . Now consider $i_1 \neq i_2$. Since $x_{i_1} \not\sim x_{i_2}$, we can find δ_w such that

$$\begin{aligned} \langle \delta_w, y \rangle &= 0, \text{ for all } y \in Y \\ \langle \delta_w, x_{i_1} \rangle &= \langle w_t, x_{i_1} \rangle \\ \langle \delta_w, x_{i_2} \rangle &= \langle w_f, x_{i_2} \rangle, \end{aligned}$$

as the $k + 2$ linear equations ($k = |Y| < n$) are independent in $n + 1$ unknowns. Now consider the weight vectors

$$\hat{w}(\lambda) = w(\lambda) + \mu \delta_w.$$

By choosing $\mu > 0$ sufficiently small we ensure that $f_{\hat{w}(1)} = f_{w(1)}$ and $f_{\hat{w}(0)} = f_{w(0)}$. Now consider $\hat{\lambda}_i$ such that

$$\langle \hat{w}(\hat{\lambda}_i), x_i \rangle = 0.$$

But $\langle \hat{w}(\lambda_{\min}), x_{i_1} \rangle = \mu \langle w_t, x_{i_1} \rangle$ so that $\hat{\lambda}_{i_1} > \lambda_{\min}$, while $\langle \hat{w}(\lambda_{\min}), x_{i_2} \rangle = \mu \langle w_f, x_{i_2} \rangle$ implying $\hat{\lambda}_{i_2} < \lambda_{\min}$. Hence we can choose a value of λ between λ_{i_1} and λ_{i_2} to obtain a function $f_{\hat{w}(\lambda)}$ which is strictly between t and f , contradicting the minimality of f . \square

Proposition 2.21 *For any sets $X, Y \subseteq \{-1\} \times \{0, 1\}^n$ with $X \cap Y = \emptyset$, let $H(X, Y)$ be the set of linearly separable functions restricted to X while shattering Y . We can bound the sum of the specification numbers of elements of $H(X, Y)$ as follows.*

$$\sum_{t \in H(X, Y)} \sigma_{H(X, Y)}(t) \leq |H(X, Y)| \log |H(X, Y)|.$$

Proof: We prove the result by induction on the number of hypotheses in the set $H(X, Y)$. For $|H(X, Y)| = 0, 1$ or 2 the result clearly holds. We first move to the input space X/Y . By Lemma 2.19 this will not affect the length of the specifying samples of elements of $H(X, Y)$ and will leave all functions distinct. Let H denote the set of functions $H(X/Y, Y)$ and choose $x \in X/Y$ such that there are two functions in H disagreeing on x . Let H' be the set of functions $H(X/Y \setminus \{x\}, Y)$ and H'' the set of functions $H(X/Y \setminus \{x\}, Y \cup \{x\})$. Then $1 \leq |H''| \leq |H'|$. For a function $f \in H$ denote by f' its restriction to $X/Y \setminus \{x\}$, which lies in H' , For $j = 0, 1$ let

$$H'_j = \{f' | f \in H \text{ and } f(x) = j\},$$

so that $H'' = H'_0 \cap H'_1$ and $H' = H'_0 \cup H'_1$. Note that $|H| = |H''| + |H'|$, since each function in $H' \setminus H''$ corresponds to exactly one function in H and each function in H'' corresponds to two (which can be distinguished only by the example x). A specifying sample for h_i in H can be constructed from the example x and a specifying sample for h' in H'_j , where $j = h_i(x)$. For hypotheses h such that $h' \in H' \setminus H''$ a specifying sample for h' in H' is also a specifying sample for h in H . Using these results we now bound the sum of the specification numbers for $h \in H$.

$$\sum_{h \in H} \sigma_H(h) \leq \sum_{t \in H''} (\sigma_{H'_0}(t) + \sigma_{H'_1}(t) + 2) + \sum_{t \in H' \setminus H''} \sigma_{H'}(t) \quad (1)$$

Below we will prove that for $t \in H''$

$$\sigma_{H'_0}(t) + \sigma_{H'_1}(t) \leq \sigma_{H''}(t) + \sigma_{H'}(t).$$

Assuming for the moment that this is true we obtain from (1) using the induction hypothesis,

$$\begin{aligned} \sum_{h \in H} \sigma_H(h) &\leq 2|H''| + \sum_{t \in H''} \sigma_{H''}(t) + \sum_{t \in H'} \sigma_{H'}(t) \\ &\leq 2|H''| + |H''| \log |H''| + |H'| \log |H'| \end{aligned}$$

Let $y = |H''|$ and $z = |H'|$. It will be sufficient to prove that $2y + y \log y + z \log z \leq (y + z) \log(y + z)$, as this will imply the required inequality

$$\sum_{h \in H} \sigma_H(h) \leq |H| \log |H|.$$

Letting $0 \leq p = y/(z + y) \leq 1/2$ after rearranging terms we obtain that the above inequality holds if and only if

$$f(p) = 2p + p \log p + (1 - p) \log(1 - p) \leq 0.$$

Since $f(0) = f(1/2) = 0$ and $f''(p) = 1/(p(1 - p)) \geq 0$ for $p \in [0, 1/2]$, the result follows.

It remains only to prove the result assumed above for $t \in H''$, namely that

$$\sigma_{H'_0}(t) + \sigma_{H'_1}(t) \leq \sigma_{H''}(t) + \sigma_{H'}(t).$$

We first show that the minimal error sets of a target $t \in H''$ in H'_i are singletons for $i = 0, 1$. This is true for H and H' by Proposition 2.20. Let $f_i \in H'_i$ be a minimal error function for t in H'_i and let f be the extension of f_i to X with $f(x) = i$. Extend t to $t_i \in H$ with $t_i(x) = i$ (this can be done since $t \in H''$). Take $g \in H$ to be a minimal error function for t_i in H , with the error set of g (with respect to t_i) a subset of the error set of f . The error set of g is a singleton subset and since f agrees with t_i on x , so must g . Hence the error set of g consists of some example not equal to x . It follows that the restriction g' of g to $X \setminus \{x\}$ has a singleton error set with respect to t , this error set being a subset of the error set of f_i . Since f_i is presumed minimal, $f_i = g'$ and so f_i has a singleton error set. Hence to specify the target in H'_i we need only the essential examples in $S_{H'_i}(t)$. Clearly

$$S_{H'_0}(t) \cup S_{H'_1}(t) \subseteq S_{H'}(t).$$

We will therefore complete the proof if we show that

$$y \in S_{H'_0}(t) \cap S_{H'_1}(t)$$

implies that y must appear in a specifying sample for $t \in H''$. But for such an example y , there exist $f_i \in H$, $i = 0, 1$, such that $z \neq x, y$ implies $f_i(z) = t(z)$, $f_i(x) = i$, and $f_i(y) \neq t(y)$. But then $f'_0 = f'_1$ determines a minimal error function in H'' with singleton error set $\{y\}$. Hence y is essential for the specification of t in H'' as required. \square

Corollary 2.22 *For the set H_n of linearly separable Boolean functions on $\{0, 1\}^n$, we can bound the sum of the specification numbers of functions in H_n as follows.*

$$\sum_{t \in H_n} \sigma_n(t) \leq |H_n| \log |H_n|,$$

for all n . \square

Proof: We can write the set of functions H_n as $H(X, \emptyset)$ where $X = \{0, 1\}^n$ and apply the proposition. \square

Since, for all n , $|H_n|$ is at most 2^{n^2} , we have the following bound on the average, or expected, specification number of a linearly separable Boolean function.

Corollary 2.23 *For the set H_n of linearly separable Boolean functions on $\{0, 1\}^n$, the average specification number $\overline{\sigma_n(t)}$ satisfies*

$$\overline{\sigma_n(t)} = \frac{1}{|H_n|} \sum_{t \in H_n} \sigma_n(t) \leq n^2$$

for all n . \square

Given that specification numbers can be exponential in n , this bound is surprisingly close to the absolute lower bound of $n + 1$. However, we cannot rule out the possibility that $\overline{\sigma_n(t)} \leq cn$ for some constant c , and it would be of interest to determine the true rate of growth of $\overline{\sigma_n(t)}$.

3 Computational Issues

Goldman and Kearns [11] raised the question of the complexity of computing the teaching dimension of a hypothesis space. We can show this is a difficult problem for a fairly simple class of hypothesis spaces. First we consider the related decision problem for specification numbers:

SPECIFICATION NUMBER

Instance A triple (H, t, k) , where H is a hypothesis space containing t and $k \leq |H|$ is an integer.

Question Is $\sigma_H(t) \leq k$?

Shinohara and Miyano [24] have shown that this problem is NP-hard by reduction to *SET HITTING* (see also Cherniavsky *et al.* [8]). We give here a proof of the NP-hardness, reducing from the well-known minimum set covering problem (see [10]).

An instance of *MINIMUM COVER* is a collection $\mathcal{S} = \{S_1, S_2, \dots, S_m\}$ of finite sets and an integer $k \leq m$. We denote by U the set $U = \bigcup_{i=1}^m S_i = \{u_1, u_2, \dots, u_n\}$. The size of such an instance may be taken to be mn . From \mathcal{S} we create an instance of *SPECIFICATION NUMBER* as follows. We take X , the example space, to be a set $X = \{x_1, x_2, \dots, x_m\}$ of m elements, and we define $H = \{h_1, h_2, \dots, h_n\} \cup \{\xi\}$ where ξ is the identically-0 function on X and, for $1 \leq i \leq n$, h_i is the $\{0, 1\}$ -valued function on X given by $h_i(x_j) = 1 \iff u_i \in S_j$ ($1 \leq j \leq m$). This instance can be constructed in polynomial time and has size $m(n+1)$. This reduction was also used in [24], where it was noted that the well-known set-covering heuristic [16] could be used to give an approximation algorithm for *SPECIFICATION NUMBER*.

Proposition 3.1 *For an integer k , \mathcal{S} has a subcovering by k of the sets if and only if $\sigma_H(\xi) \leq k$.*

Proof: We claim that the sample $\mathbf{x} = (x_{i_1}, x_{i_2}, \dots, x_{i_k})$ is a specifying sample for ξ in H if and only if the sets $S_{i_1}, S_{i_2}, \dots, S_{i_k}$ form a subcovering of the original cover — that is, if and only if their union is the whole of $U = \{u_1, u_2, \dots, u_n\}$. The result follows immediately from this claim. The claim is straightforward once we recall that the positive examples of h_i are precisely the examples x_j for j such that S_j contains u_i .

Any specifying sample for ξ must contain examples to rule out any other hypothesis in H and so it must contain a positive example of each of h_1, h_2, \dots, h_n . Thus, for each i there is $\phi(i) \in \{i_1, i_2, \dots, i_k\}$ such that $x_{\phi(i)}$ is a positive example of h_i , whence u_i belongs to the set $S_{\phi(i)}$. This shows that the collection $S_{i_1}, S_{i_2}, \dots, S_{i_k}$ covers U . The converse is similar. If this collection covers U then for each i there is $\psi(i) \in \{i_1, i_2, \dots, i_k\}$ such that u_i belongs to $S_{\psi(i)}$. Then the hypothesis h_i is ruled out by the example $x_{\psi(i)}$ in the sample. This holds for each i , so the sample specifies ξ . \square

Since *MINIMUM COVER* is NP-complete [10, 17], we have the following corollary.

Corollary 3.2 *SPECIFICATION NUMBER is NP-hard.* □

Let us now turn our attention to the problem of computing the teaching dimension of a hypothesis space:

TEACHING DIMENSION

Instance Hypothesis space H and integer $k \leq |H|$.

Question Is the teaching dimension of H at most k ?

It is known (see [10]) that *MINIMUM COVER* remains NP-complete when the sets S_i each have cardinality exactly 3. Let us denote this restricted covering problem by *X3C*. Using this result, we can prove that computing the teaching dimension is difficult for some fairly simple hypothesis spaces.

We shall say that a hypothesis space T defined on an example space X is *trivalent* if any example in X is a positive example of exactly three hypotheses in T . (Note that this is not the same as saying that each hypothesis has three positive examples, but is, in a sense, dual to this.)

Theorem 3.3 *SPECIFICATION NUMBER remains NP-hard when restricted to instances $(T \cup \{\xi\}, \xi, k)$ where T is trivalent.*

Proof: This follows directly from the reduction given above, and from the fact that *X3C* is NP-complete. Under the reduction described above, the resulting *SPECIFICATION NUMBER* problem asks whether the specification number of ξ is at most k in a hypothesis space $H = T \cup \{\xi\}$ where T is trivalent. □

Theorem 3.4 *TEACHING DIMENSION is NP-hard, and remains NP-hard when we consider only spaces of the form $H = T \cup \{\xi\}$ where T is a trivalent hypothesis space.*

Proof: Suppose that \mathcal{S} is an instance of *X3C* in which the union of the sets in \mathcal{S} has cardinality at least 9. Since each set in \mathcal{S} has cardinality 3 and the union of these sets has cardinality 9, it is clear that any subcovering consists of at least 3 sets. If H is the hypothesis space resulting from the reduction described above, then this means that $\sigma_H(\xi) \geq 3$. On the other hand, for any $t \in H$ with $t \neq \xi$, $\sigma_H(t) \leq 3$. For, if we present a positive example of t then, by the trivalent property, there are 3 hypotheses t, h, g in H which agree with t on this example. Now present a positive example of h which is a negative example of t . This rules out h (and possibly also g). If g remains, rule it out in the same way by presenting a negative example of t which is a positive example of g . Since $\sigma(\xi) \geq 3$ and $\sigma(t) \leq 3$ for $t \neq \xi$, it follows

that $\text{TD}(H) = \sigma_H(\xi)$. Thus the answer to the instance of *TEACHING DIMENSION* is the same as the answer to the instance (H, ξ) of *SPECIFICATION NUMBER*, and hence answering the *TEACHING DIMENSION* problem also answers the *MINIMUM COVER* question. The result follows immediately from the above result. \square

In summary, computing specification numbers and teaching dimensions is computationally intractable for many hypothesis spaces with some degree of structure.

We finish our discussion of complexity issues by remarking that the problem *MINIMUM UNIVERSAL SEQUENCE* (or its associated decision problem) of determining the length of a minimal universal sequence is NP-hard. This follows from the NP-completeness of the following problem (see [10]).

MINIMUM TEST SET

Instance A collection \mathcal{S} of subsets of a finite set U , and an integer $k \leq |\mathcal{S}|$.

Question Is there a subset \mathcal{S}' of \mathcal{S} of cardinality at most k with the property that for each $u, v \in U$ there is $S \in \mathcal{S}'$ which contains precisely one of u, v ?

Proposition 3.5 *MINIMUM UNIVERSAL SEQUENCE is NP-hard.*

Proof: Apply the same reduction as before, reducing from *MINIMUM TEST SET*. \square

4 Conclusions and Further Work

The main contribution of this paper is a fairly detailed study of the number of examples needed to specify exactly a linearly separable Boolean function; that is, to *teach* it to *any* consistent learner. There is an easily stated open problem directly related to the work presented here. We showed that nested hypotheses have lowest possible specification number, but the converse of this remains open: if $t \in H_n$ has specification number $n + 1$, is t necessarily a nested hypothesis?

The class of linearly separable Boolean functions is but one class of Boolean functions and it may be of interest to carry out similar analyses for other simple classes. Goldman and Kearns [11] have done this for some classes. In addition, Shinohara and Miyano [24] have obtained a simple (polynomial) upper bound on the specification numbers for the class of linearly separable Boolean functions (a subclass of H_n , in which the vector α defining the hyperplane must be a $\{0, 1\}$ -vector and the threshold θ must be a non-negative integer).

Specification is difficult because all false hypotheses must be ruled out by the sample. It would be interesting to quantify the number of examples needed to teach a hypothesis when a particular learning algorithm is being used; in this case, not all the hypotheses need to be ruled out because the algorithm may not produce them. Salzberg *et*

al. [23] have discussed this in the context of learning geometric concepts by the nearest-neighbour algorithm. Another interesting line of research is to pursue an idea of approximate specification, such as that developed by Romanik and Smith [21, 22], and to investigate the number of examples needed for approximate specification in various hypothesis spaces. Of course, both these ideas may be combined and we may ask for approximate specification by a teacher who knows the learning algorithm the learner is using. Salzberg *et al.* have results along this line when the learning algorithm is the nearest-neighbour algorithm.

There are many questions on the complexity of computing specification numbers. For example, is it NP-hard to determine the specification number of a hypothesis in H_n , the set of linearly separable Boolean functions? Shinohara and Miyano [24] have produced a polynomial-time algorithm for yielding small specifying samples in the class of Boolean threshold functions (a strict subclass of the linearly separable Boolean functions). Boros *et al.* [7] have devised a polynomial time algorithm which uses membership queries (see [2]) to learn the class of 2-monotonic positive Boolean functions (a class which includes the increasing linearly separable functions). This yields an algorithm enabling a teacher to produce small specifying samples for linearly separable Boolean functions. Are there other hypothesis spaces in which specification numbers or small specifying samples can easily be generated? These questions require further work.

Acknowledgements

We thank Dave Cohen for helpful discussions in the initial stages of this research and we thank Kathleen Romanik for comments on an early draft of part of this paper.

References

- [1] Ian Anderson, *Combinatorics of Finite Sets*, Oxford University Press, Oxford, UK, 1987.
- [2] Dana Angluin, Queries and concept learning, *Machine Learning*, 2(4), 1988: 319–342.
- [3] Martin Anthony and Norman Biggs, *Computational Learning Theory: An Introduction*, Cambridge University Press, Cambridge, UK, 1992.
- [4] Martin Anthony and John Shawe-Taylor, Using the perceptron algorithm to find consistent hypotheses. To appear, *Combinatorics, Probability and Computing*.
- [5] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler and Manfred Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *Journal of the ACM*, 36(4), 1989: 929–965.
- [6] Béla Bollobás, *Combinatorics: Set Systems, Hypergraphs, Families of Vectors and Combinatorial Probability*, Cambridge University Press, Cambridge, UK, 1986.

- [7] Endre Boros, Peter L. Hammer, Toshihide Ibaraki and Kazuhiko Kawakami, Identifying 2-monotonic positive Boolean functions in polynomial time, RUTCOR Research Report 41–91, Rutgers Center for Operations Research, Rutgers University, 1991.
- [8] J.C. Cherniavsky, R. Statman and M. Velauthapillai, Inductive inference – an abstract approach, In *Proceedings of the 1988 Workshop on Computational Learning Theory*, Morgan Kaufmann, San Mateo, CA, 1988.
- [9] Thomas M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Electronic Computers* 14, 1965: 326–334.
- [10] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.
- [11] Sally A. Goldman and Michael J. Kearns, On the complexity of teaching, In *COLT'91, Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, Morgan Kaufmann, San Mateo, CA, 1991.
- [12] Sally A. Goldman, Michael J. Kearns and Robert E. Schapire, Exact identification of circuits using fixed points of amplification functions, In *Proceedings of the Thirty-First Annual Symposium on Foundations of Computer Science*, Association for Computing Machinery Press, New York, 1990.
- [13] S.E. Hampson and D.J. Volper, Linear function neurons: structure and training. *Biological Cybernetics* 53, 1986: 203–217.
- [14] Sze-Tsen Hu, *Threshold Logic*, University of California Press, Berkeley, 1965.
- [15] Jeffrey Jackson and Andrew Tomkins, A computational model of teaching, In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Association for Computing Machinery Press, New York, 1992.
- [16] D.S. Johnson, Approximation algorithms for combinatorial problems, *Journal of Computer and Systems Sciences*, 9, 1974: 256–278.
- [17] R.M. Karp, Reducibility among combinatorial problems. In *Complexity of Computer Computations* (ed. R.E. Miller and J.W. Thatcher), Plenum Press, New York, 1972.
- [18] Nick Littlestone, Learning quickly when irrelevant attributes abound: a new linear threshold learning algorithm. *Machine Learning*, 2(4), 1988: 285–318.
- [19] C.H. Mays, Adaptive threshold logic, Technical Report 1557-1, Stanford Electronics Laboratories, Stanford University, 1963.
- [20] R. L. Rivest, Learning decision lists. *Machine Learning* 2 (3), 1987: 229–246.
- [21] Kathleen Romanik and Carl Smith, Testing geometric objects, Technical Report UMIACS-TR-90-69, CS-TR-2437, University of Maryland, Maryland, 1990.

- [22] Kathleen Romanik, Approximate testing and learnability, In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Association for Computing Machinery Press, New York, 1992.
- [23] Steven Salzberg, Arthur Delcher, David Heath and Simon Kasif, Learning with a helpful teacher, Technical Report 90/14, Computer Science, Johns Hopkins University, 1990.
- [24] Ayumi Shinohara and Satoru Miyano, Teachability in computational learning, *New Generation Computing*, 8, 1991: 337–347.
- [25] E. Sperner, Ein Satz über Untermengen einer endlichen Menge, *Math. Z.*, 27, 1928: 544–548.
- [26] V.N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2), 1971: 264-280.