Inductive Bias in Case-Based Reasoning Systems

A D Griffiths and D G Bridge

July 26, 1995

Abstract

In order to learn more about the behaviour of case-based reasoners as learning systems, we formalise a simple case-based learner as a PAC learning algorithm, using the case-based representation $\langle CB, \sigma \rangle$. We first consider a 'naive' case-based learning algorithm $CB1(\sigma_H)$ which learns by collecting all available cases into the case-base and which calculates similarity by counting the number of features on which two problem descriptions agree. We present results concerning the consistency of this learning algorithm and give some partial results regarding its sample complexity. We are able to characterise $CB1(\sigma_H)$ as a 'weak but general' learning algorithm. We then consider how the sample complexity of case-based learning can be reduced for specific classes of target concept by the application of *inductive bias*, or prior knowledge of the class of target concepts. Following recent work demonstrating how case-based learning can be improved by choosing a similarity measure appropriate to the concept being learnt, we define a second case-based learning 'algorithm' $CB2$ which learns using the best possible similarity measure that might be inferred for the chosen target concept. While $CB2$ is not an executable learning strategy (since the chosen similarity measure is defined in terms of a priori knowledge of the actual target concept) it allows us to assess in the limit the maximum possible contribution of this approach to case-based learning. Also, in addition to illustrating the role of inductive bias, the definition of $CB2$ simplifies the general problem of establishing which functions might be represented in the form $\langle CB, \sigma \rangle$. Reasoning about the case-based representation in this special case has therefore been a little more straight-forward than in the general case of $CB1(\sigma_H)$, allowing more substantial results regarding representable functions and sample complexity to be presented for $CB2$. In assessing these results, we are forced to conclude that case-based learning is not the best approach to learning the chosen concept space (the space of monomial functions). We discuss, however, how our study has demonstrated, in the context of case-based learning, the operation of concepts well known in machine learning such as inductive bias and the trade-off between computational complexity and sample complexity.

Contents

1 Introduction

This report introduces a simple model that allows the analysis of the learning behaviour of a case-based reasoning system. In essence, we apply recent formalisations of the knowledge content of a case memory system from a *functional* point of view [Jan92] [WG94] within the PAC learning model [Hau90] [Nat91] [AB92] [KV94], due originally to Valiant [Val84b]. The *functional* viewpoint sees the case-base as a representation of a mapping between input and output values. Adopting the PAC learning framework, we contrast this with the 'true' relation that holds in the application domain between input and output values, which we call the 'domain relation', and consider a *worst case, probabilistic* analysis that indicates when this mapping, with some level of confidence, will be a good approximation to the 'true' relation. In contrast with the inductive inference approach of Jantke [Jan92] and the informal approach of Wess and Globig [WG94], this 'probably approximate correct' approach has allowed us in some cases to set bounds on the number of exemplars needed to guarantee a good approximation to the domain relation. This has been our main means of comparing the different forms of case-based learning in the work reported below.

Section 2 below defines the notation to be used in this report and introduces the basic concepts of the PAC learning framework. Section 3 then describes our simple model of a case memory system which provides us with the 'case-based representation' central to this work. In sections 4 and 5, we can then consider two kinds of 'case-based learning'. In the former section (§4), we consider the simplest form of learning with the case-based representation, where 'learning' consists only of adding new cases to the case-base of the system. This is contrasted in §5 with a more elaborate case-based learning strategy, which in addition to building a case-base is able to manipulate the similarity measure by which the similarity between problem instances is calculated. The insights yielded by these studies into the nature of learning in a case memory system are given in the conclusions of §6.

2 Definitions

In this report, we deal only with the task of 'classification'; the systems being considered take a problem instance from a space of possible descriptions and output a 'yes' or 'no' answer indicating whether or not the description is a positive instance of the target concept. That is, the domain relation being modelled is a $\{0,1\}$-valued function defined on the space of possible descriptions; this is the situation that is usually assumed in the PAC learning model. The 'space of possible descriptions' will be referred to as the *example space* and denoted in the general case by X. The example space for the current work is the space of N-bit binary vectors, referred to as $D_N \mathrel{\hat{=}} \{0,1\}^N$. $(d)_i$ will be used to stand for the i-th bit of a vector $d \,\epsilon\, D_N$.

The set of $\{0,1\}$-valued total functions defined over this domain will be denoted $B_N \mathrel{\hat{=}} (D_N \to \{0,1\})$. In addition the set of *monomial* functions will specifically be referred to below. A monomial expression U is a combination of no more than N literals chosen without replacement from the set $\{u_1, \ldots, u_N\}$; additionally each chosen literal may be negated before being added to U. The classification function for the expression interprets U as a conjunction of the (possibly negated) literals:

$$h_U^N(d) = \begin{cases} 1 & \mathit{if}\ \forall i \cdot (u_i \,\epsilon\, U \to ((d)_i = 1)) \wedge (\overline{u_i} \,\epsilon\, U \to ((d)_i = 0)) \\ 0 & \mathit{otherwise} \end{cases} \tag{1}$$

The function $h_U^N(d)$ is therefore a $\{0,1\}$-valued function on D_N whose value is decided by a conjunction of the bits of d. The space of such functions will be referred to as M_N, and $M_{N,k}$ is defined as the set of monomials with exactly k literals ($\#U = k$). U_d will stand for the monomial expression corresponding to vector $d \,\epsilon\, D_N$ i.e. the unique monomial expression of size N s.t. $h_{U_d}(d') = 1 \leftrightarrow d = d'$.

In general, boolean functions will be represented by a disjunctive normal form such as $u_1u_2 + u_1\overline{u}_3$ or $u_1 + u_2\overline{u}_4u_5$ with the obvious interpretation.

Finally, we let f_0^N and f_1^N stand respectively for the 'all-0' and 'all-1' functions on D_N:

$$\forall d \in D_N \cdot f_0^N(d) = 0 \tag{2}$$
$$\forall d \in D_N \cdot f_1^N(d) = 1 \tag{3}$$

2.1 PAC Learning

For some function c from the space $X \to \{0,1\}$, the sets C^+ and C^- can be defined where $C^+ = \{x : X \mid c(x) = 1\}$ and $C^- = \{x : X \mid c(x) = 0\}$. The binary function imposes a classification on its domain, and in this situation the word 'concept' can refer to either the set of positive instances C^+ or its characteristic function c. This characteristic function c will be the 'domain relation', or in this case the *target concept* to be approximated by the learning process. In addition, the PAC framework assumes that the situation in which the learning system is deployed dictates some set of concepts which it is proposed to learn. This 'set of possible concepts' is referred to as the *concept space* C.

A second, possibly distinct set of concepts is referred to as the *hypothesis space* H. This is the space of functions that the learner can choose from in attempting to find an approximation to the target concept. The hypothesis space is determined by the *representation* used by the system to express its approximations. In many other presentations of PAC learning results, such as the results of [BEHW89], the concept space is often conflated with the hypothesis space; it is assumed that $H = C$. However, one of the important insights in this work is that the concept space and hypothesis space will often be different for a case-based learner; the distinction will be maintained carefully in what follows.

PAC-Learning applies to the situation of 'supervised learning from examples'. Here, the information used by the learner to derive its approximation to the target concept is a *training sample* of negative and positive instances of the target concept. A *sample of length* m is a member of X^m, i.e. a series of m examples. A *training* sample of length m is on the other hand an element of $(X \times \{0,1\})^m$; each of the m elements of the training sample is an example from X labelled with a 0 or a 1, designating the example as a negative or positive instance of the target concept respectively.

The framework requires that there is some fixed probability distribution μ defined over the example space X, and that each element of the samples and training samples defined above is drawn independently from X according to this probability distribution. (This means that, in general, samples will contain repeated examples.) The distribution is fixed in that the same distribution by which training examples are drawn from X is also used to measure the error in an hypothesis.

Definition 2.1 Error of an hypothesis [AB92, p.21]. *Given a target concept $t \in C$, and an hypothesis $h \in H$, where C and H are possibly distinct sets of functions, with all functions defined over the same domain X, the error of h with respect to t, denoted $er_\mu(h,t)$, is the probability that an example x drawn from X according to probability distribution μ will be misclassified.*

$$er_\mu(h,t) = \mu\{x \in X | h(x) \neq t(x)\}$$

Sufficient terms have now been defined to formalise the requirement for a learning algorithm to be *probably approximately* correct.

Definition 2.2 PAC-Learning [AB92, p.22]. *A learning algorithm L is a* probably approximately correct *learning algorithm for a concept space C if, given two constants δ and ϵ s.t.*

$$0 < \delta, \epsilon < 1,$$

then there is a positive integer m_0 which is a function of δ and ϵ only, such that for

- *all target concepts $t \in C$, and*
- *all probability distributions defined on X,*

then for any training sample s of size at least m_0, with probability $> 1 - \delta$ the error er_μ in the hypothesis inferred from that sample will be $< \epsilon$.

In other words, with *confidence* at least $1 - \delta$, the *accuracy* of the inferred hypothesis is at least $1 - \epsilon$. Thus the learner is allowed to fail, producing a bad hypothesis, for some small set of 'uninformative' samples occurring with probability $< \delta$, and additionally for all other samples a bounded amount of error ($< \epsilon$) is allowed in the hypothesis. These bounds $\delta\&\epsilon$ will however be reduced to arbitrarily small amounts by a PAC-learning algorithm if a sufficiently large training sample is presented. The size of training sample required to guarantee that a hypothesis is PAC within given values of $\delta\&\epsilon$ is referred to as the sample complexity.

Definition 2.3 Sample Complexity [AB92, p.41]. *The sample complexity m_L of a learning algorithm L with respect to a concept space C is the least value of m such that, for all concepts $t \epsilon C$, and for all probability distributions μ, the hypothesis inferred by L from a training sample of size m will, with probability $> 1 - \delta$, have an error $< \epsilon$ with respect to the target concept t.*

The sample complexity m_L will often be hard to determine exactly, so we often refer instead to an upper bound m_0 on the sample complexity, as in Definition 2.2. That is, m_0 denotes some value for the size of a training sample proven theoretically to guarantee a probably approximately correct hypothesis (for all target concepts and all probability distributions). m_0 may however be larger than the actual sample complexity due to approximations made in the proof.

3 Case-Based Representation of Classification Functions

Following the work of Jantke [Jan92], a case memory system is modelled as the pair $\langle CB, \sigma \rangle$ where CB is the case-base, or set of stored exemplars, assumed here to be free from observational error, and σ is a similarity measure defined for the space D_N. Using the terminology of Dearden's model [Dea95], the case-base is modelled as a set of pairs of 'descriptions' and 'reports'. As indicated above, a description is an N-bit vector from the space D_N. A report is a single bit denoting the classification of that exemplar, making CB an object of type:

$$CB : \mathcal{P}\ (D_N \times \{0,1\})$$

The similarity measure σ is a function over pairs of descriptions returning a normalised real value indicating the degree of similarity between the two instances:

$$\sigma : (D_N \times D_N) \rightarrow [0,1]$$

The pair $\langle CB, \sigma \rangle$ is treated as the representation of a function from B_N, according to the following interpretation related to the 'standard semantics' for a case-based classifier of Jantke and Lange [JL93]. The function represented by $\langle CB, \sigma \rangle$ is defined as:

$$h^N_{CB,\sigma}(d) = \begin{cases} 1 & if \ \ \exists (d_{pos}, 1) : CB \cdot \forall (d_{neg}, 0) : CB \cdot \sigma(d, d_{pos}) > \sigma(d, d_{neg}) \\ 0 & otherwise \end{cases} \tag{4}$$

Informally, a point d from D_N is positively classified by $h^N_{\langle CB,\sigma \rangle}$ if and only if there is a stored positive exemplar d_{pos} which is strictly more similar to d according to the chosen similarity measure σ than any of the stored negative exemplars d_{neg}. In relation to other semantics discussed by Jantke [Jan92], this interpretation resolves 'ties' between equally similar near neighbours by imposing a preference ordering on the 'report' part of retrieved cases. Negative exemplars are preferred over positive ones in inferring the classification of a new problem instance, i.e. if the set of exemplars which are most similar to d contains both positive and negative exemplars, d will be classified negatively.

In the current paper, as specific examples of a 'similarity measure', we will use either the unweighted feature count σ_H or the weighted feature count $\sigma_{\overline{w}}$. These are defined below.

Definition 3.1 'Unweighted Feature Count' Similarity Measure σ_H *The unweighted feature count* σ_H *is defined as the proportion of bits on which two members of* D_N *agree:*

$$\sigma_H(d_1, d_2) = \frac{1}{N} \sum_{i=1}^{N} (1 - |(d_1)_i - (d_2)_i|) \tag{5}$$

This measure of similarity is denoted σ_H because of its relation to the Hamming distance between the two vectors, specifically $\delta_H(d_1, d_2) = N \cdot (1 - \sigma_H(d_1, d_2))$.

Definition 3.2 'Weighted Feature Sum' Similarity Measure $\sigma_{\overline{w}}$**.** *The weighted feature sum* $\sigma_{\overline{w}}$ *is defined as the proportion of bits on which two members of* D_N *agree, where each bit of the representation has a relative importance defined by the weight vector* $\overline{w}$*:*

$$\sigma_{\overline{w}}(d_1, d_2) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \times (1 - |(d_1)_i - (d_2)_i|) \tag{6}$$

for some 'weight vector' $\overline{w} \,\epsilon\, [0,1]^N$*.*

These straightforward functions for calculating similarity share a 'weak' version of transitivity in that if descriptions d_1 and d_2 are similar according to either measure, and d_2 and d_3 are similar according to that measure, then d_1 and d_3 will also be similar. This is expressed in the results below as a modified form of the triangle inequality.

Proposition 3.1 'Transitivity' of $\sigma_{\overline{w}}$**.** *For any weight vector* $\overline{w} \,\epsilon\, [0,1]^N$*, then given similarities measured by* $\sigma_{\overline{w}}$*, if* d_1 *is similar to* d_2 *and* d_2 *is similar to* d_3*, then* d_1 *must be similar to* d_3*. Specifically:*

$$\forall \overline{w} \,\epsilon\, [0,1]^m \cdot \forall d_1, d_2, d_3 \,\epsilon\, D_N \cdot \sigma_{\overline{w}}(d_1, d_3) \geq \sigma_{\overline{w}}(d_1, d_2) + \sigma_{\overline{w}}(d_2, d_3) - 1$$

Proof: Let $\Delta_{a,b}$ be the sum of the weights of the bits on which two descriptions $d_a, d_b \,\epsilon\, D_N$ disagree; $\Delta_{a,b} = \sum_{i=1}^{N} w_i \times |((d_a)_i - (d_b)_i)|$. Clearly the triangle inequality can be applied here; $\Delta_{1,3}$ will be upper bound in terms of $\Delta_{1,2}$ and $\Delta_{2,3}$ since in the worst case $\Delta_{1,3} = \Delta_{1,2} + \Delta_{2,3}$. Hence:

$$\Delta_{1,3} \leq \Delta_{1,2} + \Delta_{2,3} \tag{7}$$

But also we have $\sigma_{\overline{w}}(d_a, d_b) = \frac{S - \Delta_{a,b}}{S}$, where $S = \sum_{i=1}^{N} w_i$. Substituting above in equation (7), this gives the stated result. □

Corollary 3.2 'Transitivity' of σ_H**.** *If* d_1 *is similar to* d_2 *and* d_2 *is similar to* d_3*, then* d_1 *must be similar to* d_3*. Specifically:*

$$\forall d_1, d_2, d_3 \,\epsilon\, D_N \cdot \sigma_H(d_1, d_3) \geq \sigma_H(d_1, d_2) + \sigma_H(d_2, d_3) - 1$$

Proof: σ_H is clearly an instance of $\sigma_{\overline{w}}$ according to a suitable choice of weight vector $\overline{w}$. Hence result follows immediately from Proposition 3.1. □

Two complementary views are presented here which provide a little leverage in the analysis of the functions represented by a case memory system. On the one hand, we consider a point in the example space to be 'labelled' with the exemplars which are most similar to that point according to the specified similarity measure. The converse notion is to consider the subset of the example space which is labelled with a particular exemplar to be the 'region of influence' of that exemplar.

Definition 3.3 Label of a point in the example space. *A point in the example space is labelled by the descriptions of the exemplars in the case-base which are most similar to it. The label is also referred to as the set of nearest neighbours* $NN(d, CB, \sigma)$*.*

$$\forall d \,\epsilon\, D_N \cdot NN(d, CB, \sigma) = \{d' \,\epsilon\, D_N | (\exists n \,\epsilon\, \{0,1\} \cdot (d', n) \,\epsilon\, CB) \wedge \forall (d'', n'') \,\epsilon\, CB \cdot \sigma(d, d') \geq \sigma(d, d'')\}$$

Proposition 3.3 *An instance $d \epsilon D_N$ is classified positively by the function $h_{\langle CB,\sigma\rangle}$ iff the label of the point d, $NN(d, CB, \sigma)$, is non-empty and contains only the descriptions of positive exemplars.*

$$\forall d \epsilon D_N \cdot h_{\langle CB,\sigma\rangle}(d) = 1 \leftrightarrow (NN(d, CB, \sigma) \neq \{\} \wedge \forall d' \epsilon NN(d, CB, \sigma) \cdot (d', 1) \epsilon CB)$$

Proof: Assuming RHS, $h_{\langle CB,\sigma\rangle}(d) = 1$ follows immediately by equation (4). Conversely, Definition 3.3 indicates that all labels will be non-empty if the case-base is non-empty. Hence an empty label indicates an empty case-base and equation (4) returns a zero on d. Alternatively, if there is a description $d' \epsilon NN(d, CB, \sigma)$ s.t. $(d', 0) \epsilon CB$, then also $h_{\langle CB,\sigma\rangle}$ is 0 on d. □

Definition 3.4 Region of Influence of an exemplar. *The region of influence of an exemplar $(d, n) \epsilon CB$ is the subset of the example space whose examples are labelled with d.*

$$RI(d, CB, \sigma) = \{x \epsilon X | \forall (d', n') \epsilon CB \cdot \sigma(x, d) \geq \sigma(x, d')\}$$

The notion of 'region of influence' allows the function represented by a case-based classifier $\langle CB, \sigma\rangle$ to be partially explicated. Equation (4) implies immediately:

$$h_{\langle CB,\sigma\rangle}(d) = 1 \leftrightarrow d \epsilon \bigcup_{(d_{pos},1) \epsilon CB} \bigcap_{(d_{neg},0) \epsilon CB} D(d_{pos}, d_{neg}, \sigma) \tag{8}$$

where $D(d_1, d_2, \sigma)$ is the subset of the example space X which is strictly more similar to the exemplar d_1 than to d_2 according to the measure σ. The expression re-expresses the claim that a point d will be positively classified if and only if there is some positive exemplar in the case-base to which d is strictly more similar than any negative exemplar, but shows that the 'regions of influence' of the positive exemplars combine independently to make the overall concept shape. This can also be expressed in terms of the boolean form of a function with a case-based representation. That is, in analogy to equation (8), the case-based representation $\langle CB, \sigma\rangle$ for a function $f \epsilon B_N$ can be 'translated' to a boolean form as follows:

$$h_{\langle CB,\sigma\rangle} = \bigvee_{(d_{pos},1) \epsilon CB} \bigwedge_{(d_{neg},0) \epsilon CB} \Delta_\sigma(d_{pos}, d_{neg}) \tag{9}$$

where $\Delta_\sigma(d_1, d_2)$ is the canonical sum containing a minterm for each point in the example space which is strictly more similar to d_1 than to d_2 according to the similarity measure σ.

Example 3.4 Interpretation of Case-Based Representation

Consider the function defined on the space D_4 by the case-base $CB = \{(1111, 1), (1001, 1), (0001, 0), (1100, 0)\}$ and the similarity measure σ_H (Definition 3.1). In the special case of σ_H, then the sum $\Delta_{\sigma_H}(d_1, d_2)$ is equivalent to the boolean function which is true if a description d agrees with d_1 on a strict majority of the bits on which d_1 and d_2 differ.

$$\Delta_{\sigma_H}(d_{pos}, d_{neg}) = \bigvee\{U \subseteq (U_{d_{pos}} \setminus U_{d_{neg}}) | \#U = \left\lfloor \frac{\#(U_{d_{pos}} \setminus U_{d_{neg}})}{2} \right\rfloor + 1\} \tag{10}$$

where $U_{d_{pos}}$ and $U_{d_{neg}}$ are the monomial expressions corresponding to the descriptions d_{pos} and d_{neg} as above. Hence the boolean representation of $f = h_{\langle CB,\sigma_H\rangle}$ is:

$$\begin{aligned}
&h_{\langle CB,\sigma_H\rangle} \\
&= \Delta(1111, 0001).\Delta(1111, 1100) + \Delta(1001, 0001).\Delta(1001, 1100) \\
&= (u_1u_2 + u_1u_3 + u_2u_3).u_3u_4 + u_1.\overline{u}_2u_4 \\
&= u_1u_3u_4 + u_2u_3u_4 + u_1\overline{u}_2u_4 \quad \square
\end{aligned}$$

It is clear that, in contrast with other symbolic AI approaches to machine learning, $\langle CB, \sigma \rangle$ is an *implicit* representation of the system's state [WG94, p.79]. That is, it is difficult to understand by inspection of the pair precisely what function is currently represented. A little experimentation with the representation shows that large changes to the boolean form of a function can result from small changes to the representing case-base and vice-versa. This makes it difficult, amongst other things, to establish whether there is a case-base containing exemplars from some target concept or concept class which gives a case-based representation of some particular function. However, Proposition 3.5 below states one fact that is easily established and has proved a useful constraint in reasoning whether specific functions may or may not have a case-based representation. Given a point in the example space which is positively classified by a function with a case-based representation, then by equation (4) there must be some positive exemplars labelling the example which are responsible for the positive classification, and the result below shows that all examples in the subset of the example space lying on the set of shortest paths between the positive instance and the exemplars influencing it must also be classified positively.

Proposition 3.5 *Consider a case-base CB and a point $d \,\epsilon\, D_N$ which is a positive instance of $h_{\langle CB, \sigma_H \rangle}$. Let d_1^+ be any positive exemplar in CB which is maximally similar to d w.r.t. σ_H:*

$$\forall (d', n) \,\epsilon\, CB \cdot \sigma_H(d, d_1^+) \geq \sigma_H(d, d')$$

Then, for any such positive exemplar, any instance which lies on a shortest path through the example space between d and d_1^+, i.e. any point $d^* \,\epsilon\, D_N$ s.t.

$$\forall 1 \leq i \leq N \cdot (d^*)_i \neq (d_1^+)_i \rightarrow (d)_i \neq (d_1^+)_i$$

is also a positive instance of $h_{\langle CB, \sigma_H \rangle}$:

$$h_{\langle CB, \sigma_H \rangle}(d^*) = 1$$

Proof: Given $h_{\langle CB, \sigma_H \rangle}(d) = 1$, then the set of nearest neighbours of d must be non-empty and contain only positive exemplars (Proposition 3.3). These exemplars will be equally similar to each other, will be maximally similar to d w.r.t. all other exemplars in the case-base and are strictly more similar to d than any negative exemplar. For any such positive exemplar d_1^+:

$$\forall (d^-, 0) \,\epsilon\, CB \cdot \sigma_H(d, d_1^+) > \sigma_H(d, d^-) \tag{11}$$

For a proof by contradiction, assume that for one such positive exemplar there is a description d^* satisfying $(d^*)_i \neq (d_1^+)_i \rightarrow (d)_i \neq (d_1^+)_i$, s.t. there is some negative exemplar $(d_1^-, 0) \,\epsilon\, CB$ which is at least as similar to d^* as d_1^+ is:

$$\exists (d_1^-, 0) \,\epsilon\, CB \cdot \sigma_H(d^*, d_1^-) \geq \sigma_H(d^*, d_1^+) \tag{12}$$

Assumed above, we have:

$$\forall 1 \leq i \leq N \cdot (d^*)_i \neq (d_1^+)_i \rightarrow (d)_i \neq (d_1^+)_i \tag{13}$$

Hence, all bits on which d and d_1^+ agree, d^* and d_1^+ also agree on; clearly d and d^* must also agree in this case. Thus the number of bits on which d and d_1^+ agree is equal to the number of bits on which d and d^* agree, minus the number of bits on which d and d^* agree, but on which d^* and d_1^+ disagree. Since in addition, equation (13) requires $(d^*)_i \neq (d_1^+)_i \rightarrow (d)_i = (d^*)_i$, then all bits on which d^* and d_1^+ disagree are also agreed on by d and d^*. Thus

$$N.\sigma_H(d, d_1^+) = N.\sigma_H(d, d^*) - (N - N.\sigma_H(d^*, d_1^+))$$

and $\sigma_H(d, d_1^+) = \sigma_H(d, d^*) + \sigma(d^*, d_1^+) - 1$. Thus from equation (12):

$$\sigma_H(d^*, d_1^-) \geq \sigma_H(d, d_1^+) - \sigma_H(d, d^*) + 1 \tag{14}$$

In turn, from corollary 3.2 we have $\sigma_H(d, d_1^-) \geq \sigma_H(d, d^*) + \sigma_H(d^*, d_1^-) - 1$ Hence equation (14) gives $\sigma_H(d, d_1^-) - \sigma_H(d, d^*) + 1 \geq \sigma_H(d, d_1^+) - \sigma_H(d, d^*) + 1$, and thus from equation (12) we have derived

$$\exists d_1^+ \,\epsilon\, NN(d, CB, \sigma) \cdot \exists (d_1^-, 0) \,\epsilon\, CB \cdot \sigma_H(d, d_1^-) \geq \sigma_H(d, d_1^+) \tag{15}$$

which contradicts equation (11). Hence

$$\forall (d^-, 0) \,\epsilon\, CB \cdot \sigma_H(d^*, d^-) < \sigma_H(d^*, d_1^+) \tag{16}$$

and we conclude $h_{\langle CB, \sigma_H \rangle}(d^*) = 1$ for any d^* lying on a direct path between d and d_1^+ where d_1^+ is any positive exemplar labelling d. □

4 Case-Based Learning with Fixed Similarity Measures

Aha *et al* [AKA91, p.40] suggest that a particular case-based system of the kind discussed here is defined by the specification of a similarity function, classification function (c.f. our 'semantics') and concept description update function. Rather than formalise the updating of the system's hypothesis as a function, we follow the conventions of computational learning theory and define 'case-based learning algorithms' that describe how the system's approximation to the target concept is updated as new elements of the training sample are read. Wess and Globig emphasise that within a model of this kind, the knowledge content of the system rests in both the set of exemplars and in the chosen measure of similarity. Therefore "there are three possibilities to improve a case-based system:

- store new cases in the case base CB
- change the measure of similarity [σ]
- change CB and [σ]" [WG94, p.79]

Many 'case-based learning' algorithms have been defined illustrating these options; IB2 [AKA91], VS-CBR [WG94] and PEBLS [CS93] [YJL94] show a number of options for adjusting the represented hypothesis. The current section will study the situation where concepts are learnt using a single fixed similarity measure, and the hypothesis is updated by alterations to the case-base alone. Specifically, having defined a simple case-based learning algorithm, we will firstly (§4.1) consider how the choice of similarity measure affects the *consistency* of $CB1(\sigma)$. That is, we will give results characterising those functions σ which ensure that an algorithm $CB1(\sigma)$ has the basic property of being able to reproduce the classifications of those examples seen in the training sample. In the following sections we will then report some simple empirical observations of $CB1$ (§4.2) and our results in trying to account for these observations within PAC learning theory (§4.3). We must first, therefore, define the following family of learning algorithms.

Definition 4.1 $CB1(\sigma)$ **Learning Algorithm for Case-Based Classifiers**

```
set $CB = \emptyset$
for i = 1 to m do
    set $CB = CB \cup \{(d_i, b_i)\}$
set $CB1(\sigma)(\overline{s}) = h_{\langle CB, \sigma \rangle}$
```

where $\overline{s} = \langle (d_i, b_i) \rangle_{i=1}^m$ *is a training sample from* $(D_N \times \{0, 1\})^m$.

We refer to this as a 'family' of learning algorithms since there are clearly different instantiations of definition 4.1 which will output different hypotheses on the same training sample according to the (a priori) choice of similarity measure σ, such as the measures of similarity given in Definitions 3.1 and 3.2 above.

The algorithms designated $CB1(\sigma)$ learn by adding each and every member of the training sample $\overline{s}$ (a series of m pre-classified examples (d_i, b_i)) to the case-base. In contrast with algorithms such as IB2 [AKA91] and other 'instance filtering methods', which are deliberately more economical about the exemplars they retain [Cam92] [Zha92] [Bib95], this has the result that for a given target concept t, all possible case-bases $CB \subseteq t$ are reachable by the learning algorithm. This is expressed in the following result, which also introduces the notation $H_C^{CB1(\sigma)}$ to stand for the set of possible output hypotheses or *hypothesis space* of $CB1(\sigma)$ with respect to some concept space C.

Proposition 4.1 *A function f is a member of the hypothesis space of $CB1(\sigma)$ with respect to the concept space $C \subseteq B_N$ if and only if there is some target concept $c \in C$ for which there is a case base $CB \subseteq c$ s.t. $h_{\langle CB,\sigma\rangle} = f$.*

$$\forall C \subseteq B_N \cdot \forall f \in B_N \cdot f \in H_C^{CB1(\sigma)} \leftrightarrow \exists c \in C \cdot \exists CB \subseteq c \cdot h_{\langle CB,\sigma\rangle} = f$$

Proof: **a) Only if.** Assume that a function f is output as a hypothesis from the algorithm on $\overline{s}$, which is a training sample for the target concept t. By Definition 4.1, f is a function $h_{\langle CB,\sigma_H\rangle}$, where CB contains exactly those examples presented in the training sample. Since we have assumed that there is no observational error or noise of any kind, then for any (d_i, b_i) in the sample, $t(d_i) = b_i$, and hence $CB \subseteq t$. Hence RHS. **b) If.** Assume that a function f has a case-based representation $\langle CB, \sigma\rangle$, where $CB \subseteq c$ for some $c \in C$. Clearly, any enumeration of the exemplars in CB can be presented to $CB1(\sigma)$ as a training sample on which the algorithm outputs the hypothesis f. Hence LHS. □

As a corollary of Proposition 4.1:

$$H_C^{CB1(\sigma)} = \bigcup_{t \in C} hyp_{t,\sigma} \tag{17}$$

where $hyp_{t,\sigma} = \{h_{\langle CB,\sigma\rangle} | CB \subseteq t\}$

These definitions show how the hypothesis space of $CB1(\sigma)$ depends on the choice of both the similarity measure σ and the concept space C. This gives the analysis of case-based learning algorithms a distinctive feel in contrast to the way that in most learning algorithms a fixed hypothesis space is determined by the 'hard-wired' representation used by the learner. Note also that for smaller concept spaces, it will not be uncommon that a case base CB which is extensible to some target concept c will be interpreted by equation (4) as a function from outside of the concept space. This is illustrated in the following example which makes use of the unweighted feature count σ_H.

Example 4.2 Hypothesis Space of $CB1(\sigma_H)$ with respect to the concept space $M_{N,1}$

The concept space $M_{N,1}$ (see §2) is the space of functions which partition D_N into two equal halves according to the value of a single bit of the representation, i.e. the set of monomial functions which can be represented by a single literal.

Consider the function $h_{\{u_1\}}$ from $M_{N,1}$, and training sample $\overline{s} = \langle(1111, 1), (0000, 0)\rangle$ for $h_{\{u_1\}}$. Note that the first exemplar is a positive instance since the first bit is 1, while the second exemplar is a negative instance. $CB1(\sigma_H)$ will convert $\overline{s}$ to the case-base $CB' = \{(1111, 1), (0000, 0)\}$, and infer $h_{\langle CB',\sigma_H\rangle}$ as an approximation to $h_{\{u_1\}}$. Following Example 3.4, $h_{\langle CB',\sigma_H\rangle} = \Delta_{\sigma_H}(1111, 0000) = u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_4 + u_2u_3u_4$. Hence $h_{\langle CB',\sigma_H\rangle} \notin M_{N,1}$. □

The nature of this relationship will be illustrated further below, and, in particular, the impact that it has on the *sample complexity* of $CB1(\sigma)$ as a learning algorithm. First, however, it is necessary to consider the property of *consistency* of a case-based learning algorithm.

4.1 Consistency of $CB1(\sigma)$

Definition 4.2 Consistency of a learning algorithm. *A learning algorithm L is consistent with respect to a class of functions C iff for all training samples s for any target concept $t \in C$, the hypothesis $h = L(s)$ inferred from that sample correctly classifies all examples present in the training sample.*

$$\forall 1 \leq i \leq m \cdot h(x_i) = t(x_i)$$

where $\overline{s} = \langle (x_1, t(x_1)), \ldots, (x_n, t(x_n)) \rangle$.

Consistency is one of the most basic properties of a learning algorithm, being the property that the algorithm will produce a hypothesis which will classify correctly at least the examples in its training sample. It is not always desirable for an algorithm to be consistent. This is certainly the case in noisy domains; the *noise tolerant* learner IB3 [AKA91] attempts to disagree with the training sample on precisely those instances which are suspected to be noisy, and Turney [Tur93] gives a formalisation of the sense in which a consistent hypothesis which always agrees with the training sample may be sub-optimal in the presence of noise. The current work assumes however that the domain is free from noise, and therefore can make use of results applying specifically to the class of consistent learners. It is clear that this class will have certain regularities of behaviour which will facilitate their analysis, and there are indeed a number of standard results bounding the sample complexity (Definition 2.3) of learning algorithms which hold in general for consistent learning algorithms.

The main result of this section, Theorem 4.3, gives necessary *and* sufficient conditions over σ to make $CB1(\sigma)$ a consistent learning algorithm. This will be a valuable tool in making use of the sample complexity results for consistent learning algorithms. Results elsewhere [JL93, Lemma 3] [Tur93, Lemma 7] formalise the intuition that a 'reasonable' similarity measure [Tur93], which recognises that an object is more similar to itself than any other object, will be sufficient for consistency. This property is here called 'definiteness' after Day and Faith [DF86, p.183].

Definition 4.3 Definiteness of a Similarity Measure. *A similarity measure σ is definite iff the comparison of two distinct objects yields a score strictly less than the score given to the comparison of an object to itself.*

$$\forall d, d' : D_N \cdot d \neq d' \rightarrow \sigma(d, d') < \sigma(d, d) \tag{18}$$

This property ensures a consistent hypothesis since any exemplar in the case base will be judged strictly most similar to itself, and therefore those exemplars at least will be classified correctly by equation (4). Definiteness is not however a necessary condition for consistency. The exemplars in the case base will still be classified correctly as long as the most similar object to a positive exemplar is *any* positive exemplar and the most similar object to a negative exemplar is *any* negative one. In other words, two distinct objects may be assigned maximal similarity only if they are classified the same by all relevant classification functions f. This is recognised informally as a necessary condition by Wess and Globig [WG94, p.86]. We express it within our framework in our definition of predictivity and prove it a necessary and sufficient condition over σ to make $CB1(\sigma)$ a consistent learning algorithm.

Definition 4.4 Predictivity of a Similarity Measure with respect to a concept space C. *A similarity measure is predictive of a concept space C iff, for any concept $c \in C$:*

1. *When d is a positive instance of c, the comparison of d and d' yields a score at least as large as the comparison of d to itself only if d' is also a positive instance.*

$$\forall c \in C \cdot \forall d, d' \in D_N \cdot \sigma(d, d') \geq \sigma(d, d) \rightarrow c(d) = 1 \rightarrow c(d') = 1 \tag{19}$$

2. *When d is a negative instance of c, the comparison of d and d' yields a score strictly greater than the comparison of d to itself only if d' is also a negative instance.*

$$\forall c \in C \cdot \forall d, d' \in D_N \cdot \sigma(d, d') > \sigma(d, d) \rightarrow c(d) = 0 \rightarrow c(d') = 0 \tag{20}$$

Note how this relates to equation (4) in that the property of definiteness is relaxed precisely where no misclassification will occur under our chosen classification function (4). The asymmetry in equations (19) and (20) reflects the preference given to negative exemplars in the classification function. Hence we emphasise that choosing a different semantics in (4) would entail a slightly different form of the following theorem.

Theorem 4.3 Consistency of $CB1(\sigma)$. *For any concept space $C \subseteq B_N$, $CB1(\sigma)$ is a consistent learning algorithm for C if and only if the chosen similarity measure σ is predictive of C.*

Proof: Let $CB1(\sigma)$ infer a hypothesis from some training sample $\overline{s} = \langle (d_i, b_i) \rangle$ for a target concept c. According to the definition of $CB1(\sigma)$, the case-base will contain exactly those labelled examples presented in the training sample; in the absence of observational error we can assume $(x, n) \,\epsilon\, CB \rightarrow c(x) = n$, for $n \,\epsilon\, \{0, 1\}$. **a) Sufficiency:** Assume σ is predictive of the concept space C. Taking positive and negative exemplars in the case base separately, consider first d_i such that $b_i = 1$. For any negative exemplar $(d_{neg}, 0) \,\epsilon\, CB$, we have $c(d_{neg}) = 0$ and hence by equation (19) $\forall d \,\epsilon\, D_N \cdot \forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma(d, d_{neg}) < \sigma(d, d) \vee c(d) = 0$. Since $c(d_i) = 1$, we conclude $\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma(d_i, d_{neg}) < \sigma(d_i, d_i)$, and thus $h_{\langle CB, \sigma \rangle}(d_i) = 1$ by equation (4). By a similar argument, for some d_i such that $b_i = 0$ we derive from equation (20), $\forall (d_{pos}, 1) \,\epsilon\, CB \cdot \sigma(d_i, d_{pos}) \leq \sigma(d_i, d_i)$, and hence $h_{\langle CB, \sigma \rangle}(d_i) = 0$ by equation (4). Thus for any example d_i in an arbitrary training sample, $h_{\langle CB, \sigma \rangle}(d_i) = b_i$, making $CB1(\sigma)$ a consistent learning algorithm. **b) Necessity.** It will be shown that for any similarity measure σ' which violates either of equations (19) & (20), there is a target concept c' from the specified concept space for which a training sample can be constructed which will be mis-classified by $CB1(\sigma)$. The consistency of $CB1(\sigma)$ would therefore require a similarity measure satisfying both equations. If equation (19) does not hold, then there must be two descriptors d_1 and d_2 and a target concept c' such that:

$$\sigma'(d_1, d_2) \geq \sigma'(d_1, d_1) \wedge c'(d_1) = 1 \wedge c'(d_2) = 0 \tag{21}$$

Thus $\langle (d_1, 1), (d_2, 0) \rangle$ will be a training sample for c'. Given the case base CB constructed by $CB1(\sigma')$ from this sample, note that $h_{CB, \sigma'}(d_1) = 0$ since equation (21) indicates that the negative exemplar d_2 will be at least as similar to d_1 as d_1 is to itself. Hence $h_{\langle CB, \sigma \rangle}$ disagrees with the training sample. In a similar way, if it assumed that equation (20) is relaxed, then there is a training sample $\langle (d_1, 0), (d_2, 1) \rangle$ resulting in a hypothesis such that $h_{\langle CB, \sigma' \rangle}(d_1) = 1$. Thus $CB1(\sigma)$ will be a consistent learning algorithm for a concept space C if and only if σ is predictive of C. □

The close relationship between Definition 4.3 (definiteness) and Definition 4.4 (predictivity) means that the following additional result can be easily established:

Corollary 4.4 *$CB1(\sigma)$ is a consistent learning algorithm for the space B_N of all total functions on D_N if and only if σ is a definite similarity measure.*

Proof: **a) Sufficiency.** Definition 4.3 entails both (19) and (20). Hence result by Theorem 4.3. **b) Necessity.** Assume $CB1(\sigma)$ is a consistent learning algorithm for B_N. Hence, by Theorem 4.3:

$$\forall d, d' \,\epsilon\, D_N \cdot \sigma(d, d') \geq \sigma(d, d) \rightarrow \forall f \,\epsilon\, B_N \cdot f(d) = 1 \rightarrow f(d') = 1 \tag{22}$$

Assume also two distinct exemplars d_1 and d_2. B_N contains every possible dichotomy of D_N, and so $\exists f \,\epsilon\, B_N \cdot f(d_1) = 1 \wedge f(d_2) = 0$. Equation 22 gives $\sigma(d_1, d_2) < \sigma(d_1, d_1)$, hence we conclude that σ must be a definite similarity measure (Definition 4.3). □

Having established the precise conditions under which $CB1(\sigma)$ is a consistent learner, the following result follows trivially.

Corollary 4.5 *A similarity measure σ which is predictive of a concept space C is sufficient to ensure that $CB1(\sigma)$ is a PAC learning algorithm for C.*

Proof: Any learning algorithm which is consistent with respect to some concept space and which learns using a *finite* hypothesis space is a PAC-learning algorithm for that concept space [AB92, p.41]. Hence result follows from Theorem 4.3 since the number of distinct binary functions that can be defined on D_N is 2^{2^N} indicating that the hypothesis space of $CB1(\sigma)$ must be finite. □

The PAC-Learnability results given for case-based classifiers in [AKA91] and [AA91] hold for concepts defined on real valued attributes. As a result of dealing with an uncountable example space, Albert and Aha have to modify the PAC learning framework by introducing constraints on the probability distribution on the example space before PAC-learnability can be proven. The finite example spaces considered in the current paper mean however that any consistent learning algorithm (and some inconsistent ones) will satisfy the 'classical' definition of PAC-learnability (Definition 2.2) without additional constraints.

PAC learnability is an important basic result because it establishes that, given enough samples, the learning algorithm will eventually converge to arbitrarily good approximations to the target concept. In addition, two other results are presented as corollaries to the property of consistency:

Corollary 4.6 *Given a similarity measure σ which is predictive of a concept space C, then for any target concept $c \epsilon C$ there is a case-base CB s.t. $h_{\langle CB,\sigma\rangle} = c$.*

Proof: For some σ and C s.t. σ is predictive of C, take any $c \epsilon C$ and any training sample $\overline{s}$ for c which contains an exemplar for every point in the example space D_N. Since Theorem 4.3 guarantees that the output of $CB1(\sigma)$ will be consistent with $\overline{s}$, clearly the function $h_{\langle CB,\sigma\rangle}$ output by $CB1(\sigma)$ on $\overline{s}$ will be exactly c. □

Corollary 4.6 is of significance in establishing the *universality* of the case-based representation; it makes it clear that any total function $f \epsilon B_N$ has at least one case-based representation. The same line of reasoning establishes the following result, which makes an initial contribution to our understanding of the effective hypothesis space of a case-based learning algorithm with respect to the choice of concept space.

Corollary 4.7 *A similarity measure σ which is predictive of a concept space C will ensure that C is contained in $H_C^{CB1(\sigma)}$, the hypothesis space of $CB1(\sigma)$ with respect to that concept space C.*

$$\forall C \subseteq B_N \cdot \mathit{Predictive}_C(\sigma) \rightarrow C \subseteq H_C^{CB1(\sigma)}$$

4.2 Empirical Investigation of $CB1(\sigma)$

Having established sufficient conditions under which we know that the basic case-based learning algorithm $CB1(\sigma)$ will *eventually* converge to a good approximation of the target concept, we now wish to consider what can be said about *how quickly* the algorithm converges to its target concept. We wish to compare the ease with which $CB1(\sigma)$ learns functions from this general class of boolean functions B_N with its performance over more restricted classes of 'easier' concepts in order to explore the factors which make a concept 'easy' or 'difficult' to learn. 'Ease of learning' is formalised by considering the size of the training sample needed to reach certain levels of confidence and accuracy for $CB1(\sigma)$ (the sample complexity - Definition 2.3). Empirical results will be presented which will motivate further discussion, and illustrate the learning capabilities of $CB1(\sigma)$.

Experimental software has been implemented to measure the learning curve for $CB1(\sigma)$ on training samples of increasing size. The probability distribution on the example space, here and in all empirical work described in this paper, is uniform. Additionally, in this section all empirical measurements are with respect to the specific learning algorithm $CB1(\sigma_H)$ instantiated with the 'unweighted feature count' similarity measure σ_H defined above (Definition 3.1) unless otherwise stated.

The experimental method used is as follows. A function is chosen randomly from the concept space, and a training sample of a fixed maximum size is generated for that concept. Each member of the training

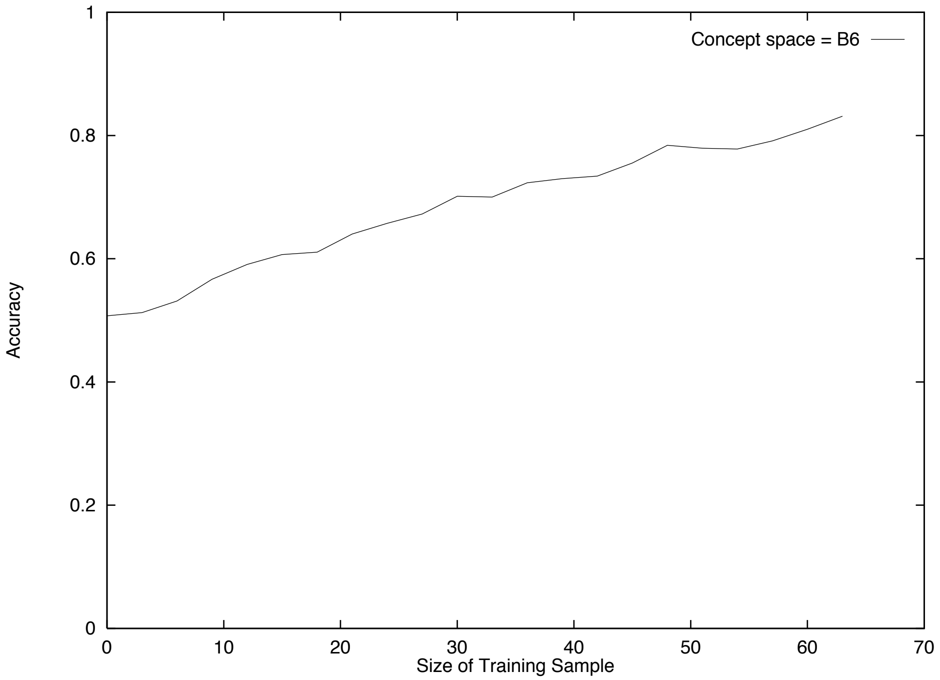

Figure 1: Average Learning Curve For All Binary Functions

sample in turn is presented to the case-base as a test instance, the correctness of the classification noted and then the same instance is added to the case memory as specified by $CB1(\sigma_H)$. (This corresponds to the interleaved testing/learning algorithm IB1 described by Aha *et al* [AKA91, p.42].) This is repeated for a large number of different training concepts and the proportion of correctly classified training instances is calculated for each value of M (the number of instances taken so far from the training sample).

Figure 1 illustrates the learning curve produced by this method for learning arbitrary functions from B_6, averaging the learning behaviour over 500 different training samples. The graph shows how the average accuracy of the hypothesis increases from roughly half for an empty case-base ($M = 0$) to $\sim 80\%$ for training samples of size 60. Additionally, Figure 2 shows a measurement of the size of the case-base (the number of distinct exemplars in the training sample) against the sample size for the same experiments. This illustrates the comment made above that the number of repetitions in the training sample will become large as the size of the training sample increases. For example, it appears that for a training sample of size 60, on average < 40 distinct exemplars will have been added to the case-base.

The above results suggest that there are 'hard' domain functions in B_N that are not suited to representation by a case-based classifier. Figures 1 and 2 indicate that, to reach even 80 % accuracy and confidence, on average over half the possible examples have to be represented in the case-base. This would not seem to be an economic use of case-based reasoning.

In order to explore what might make a concept space 'easier' for $CB1(\sigma_H)$ to learn, the above results for B_N were compared to those for a highly restricted set of binary functions. The concept space for the following experiments was therefore restricted to the set M_N of monomial functions defined in §2. The experimental method described above was reproduced to estimate the average-case accuracy of the case-based classifier for 500 randomly chosen instances of M_N for $N = 6$. Figure 3 graphs these results, showing the improved accuracy of the system compared to the previous results for B_6. The generally very high accuracy is due to the relatively small proportion of positive instances of a monomial function in the general case. Note, however, how slowly the classification accuracy improves as extra exemplars

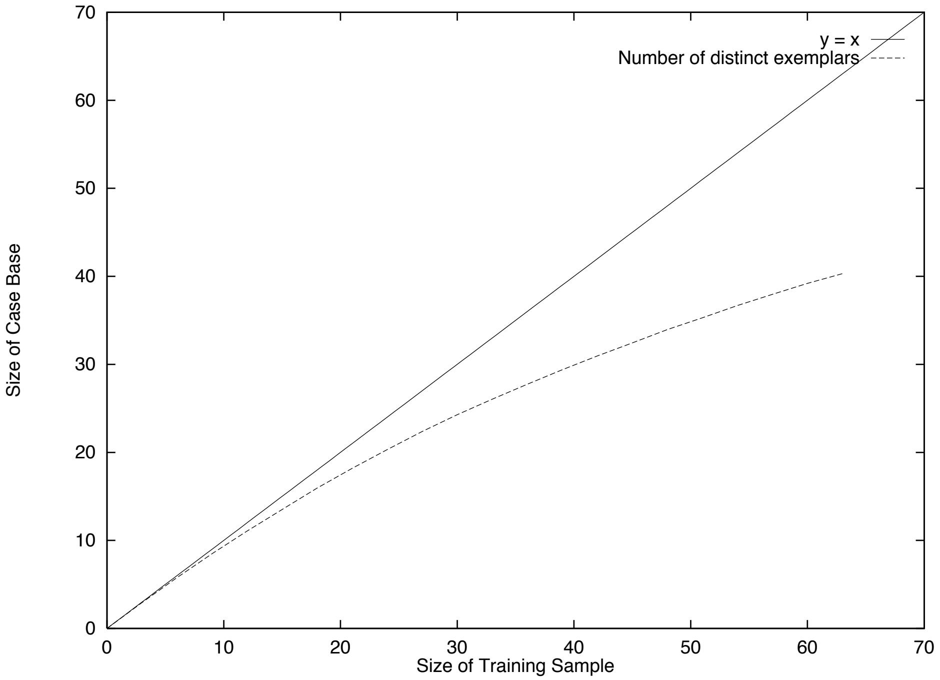

Figure 2: Number Of Distinct Instances In Training Sample

are added to the case-base.

Finally, the learning behaviour of $CB1(\sigma_H)$ over the even more restricted concept spaces $M_{N,k}$ was measured for the range of k, $1 \leq k \leq N$. We wish to study this case from the point of view of inductive bias, since Langley & Iba's meticulous average case analysis of a 'nearest neighbour algorithm for learning conjunctive concepts' [LI93] corresponds exactly to the case of learning $M_{N,k}$ by $CB1(\sigma_H)$.

The previous experiments were repeated for 250 instances of $M_{N,k}$ for values of k between 1 & 3, and for $N = 6$. Results from these experiments are shown in Figure 4. Note particularly how functions from $M_{6,1}$ are learnt more slowly than functions from $M_{6,2}$, which are learnt more slowly than functions from $M_{6,3}$.

In summary, the learning behaviour of $CB1(\sigma_H)$ is observed to improve as the concept space for the algorithm is reduced from B_6 (Figure 1) to the much more constrained space M_6 (Figure 3). Taking a partition of M_6 according to the number of 'relevant bits' defining the target concept, the space $M_{6,k}$ becomes apparently 'harder' to learn as k decreases from 3 to 1 (Figure 4). Comparing the learning curves for M_6 and $M_{6,1}$, the hypotheses produced by $CB1(\sigma_H)$ on training samples for targets in $M_{6,1}$ have apparently higher error than the average for concepts in M_6 over the range observed. This suggests that $M_{6,1}$ contains some of the 'harder' functions in M_6.

4.3 Inductive Bias in $CB1(\sigma)$

Having observed empirically some different aspects of the learning behaviour of the algorithm $CB1(\sigma_H)$, we now consider to what extent these results might have been predicted analytically. Definition 2.3, the *sample complexity* of a learning algorithm, gives a worst case size of training sample needed for the algorithm to achieve an accurate hypothesis with some degree of confidence. The following equation, known as the *Blumer Bound* [BEHW87, Lemma 2.1] [Hau88, Lemma 2.2], gives one upper bound for the sample complexity of *any* consistent learning algorithm using a finite hypothesis space, in terms of

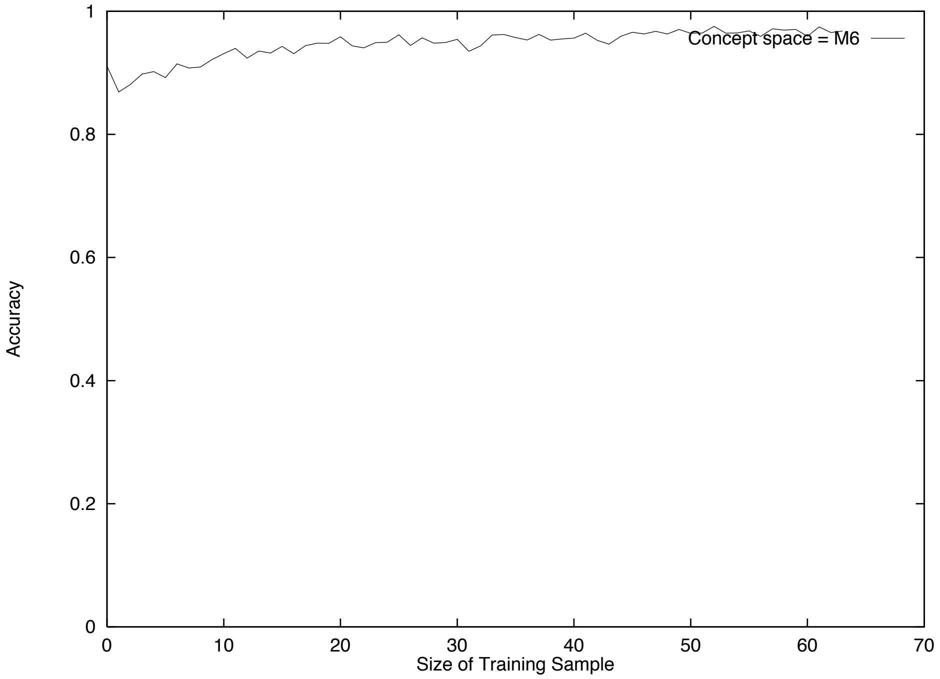

Figure 3: Average Learning Curve For Monomial Functions

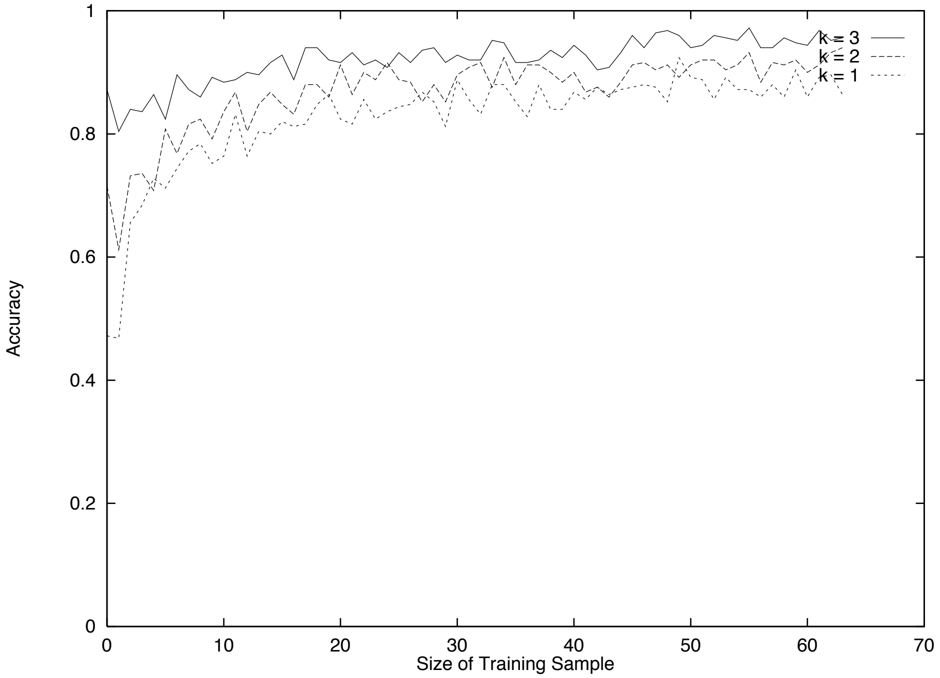

Figure 4: Average Learning Curve For Monomial Functions of k Literals by CB1(s)

the cardinality of that hypothesis space:

$$m_0(\delta, \epsilon) \leq O\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{\log|H|}{\epsilon}\right) \tag{23}$$

Alternatively, an upper bound on sample complexity may be expressed in terms of the *VC dimension* of the hypothesis space in place of the term $\log|H|$.

Definition 4.5 Behaviours of S**:** $\Pi_H(S)$ *[AB92, p.73] The set of behaviours of* S *realised by a function class* H*, denoted* $\Pi_H(S)$*, is the set of subsets* $S' \subseteq S$ *s.t. there is some function* $h \,\epsilon\, H$ *whose set of positive instances intersects with* S *to give* S'*.*

$$\Pi_H(S) = \{S' \subseteq S | \exists h \,\epsilon\, H \cdot \forall s \,\epsilon\, S \cdot h(s) = 1 \leftrightarrow s \,\epsilon\, S'\}$$

Definition 4.6 Growth function $\Pi_H(m)$ *[AB92, p.73]. The value of the growth function* $\Pi_H(m)$ *for a sample size* m *with respect to some space of functions* $H \subseteq (X \to \{0,1\})$ *is the maximum number of behaviours which can be induced on a sample of size* m *from the set* X^m *by functions from the space* H*.*

$$\Pi_H(m) = \max\{|\Pi_H(S)| \cdot |S| = m\}$$

Definition 4.7 Shattering *[AB92, p.74]. A sample of size* m *is shattered by a space of functions* H *if and only if there is a function in* H *giving each possible classification of the sample i.e.* $\Pi_H(m) = 2^m$*.*

Definition 4.8 VC Dimension *[AB92, p.74]. The VC Dimension of a space of functions* H *is the maximum value of* m *such that some sample of size* m *is shattered by* H*.*

$$d_{VC}(H) = \max\{m | \Pi_H(m) = 2^m\}$$

As an alternative to the 'Blumer Bound', the following result also applying to *any* consistent learning algorithm, gives a bound on sample complexity in terms of this 'VC dimension' [BEHW89, Thm 2.1(ii)(a)] [Hau88, Thm 4.4]. Note that in general, $d_{VC}(H) \leq \log_2|H|$ [Nat91, Lemma 2.1] and that $d_{VC}(H)$ and $\log|H|$ will often be quantities of the same order. Blumer et al do however note exceptions to this correlation [BEHW89, p. 938].

$$m_0(\delta, \epsilon) \leq O\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d_{VC}(H)}{\epsilon}\log\frac{1}{\epsilon}\right) \tag{24}$$

Sample Complexity for B_N

Equations (23) and (24) are given in terms of the cardinality of the hypothesis space (23) and its VC dimension (24). Characterising the hypothesis space of a machine learning algorithm is most easily done when the representation directly encodes the anticipated concept space, i.e. when we consider the concept space to be precisely those functions which might be represented as hypotheses. In contrast, we consider $CB1$ as an example of a learner using a general representation ($\langle CB, \sigma\rangle$) capable of representing any boolean function, but which is trained on a target concept from some restricted concept space so that only some subset of those representable functions are output as hypotheses. Clearly, we have not yet been able to give any characterisation of the effective hypothesis space of $CB1(\sigma_H)$ with respect the concept spaces considered in §4.2, and so in general are not yet able to apply the bounds of equations (23) and (24) to these cases.

The one case where equation (23) *can* be applied is in the learning of B_N by $CB1(\sigma)$ in the case of a definite similarity measure σ. Since $CB1(\sigma)$ will be a consistent learning algorithm, all 2^{2^N} functions of B_N will be included in the hypothesis space (Corollary 4.7), and (since equation (4) interprets only as a total function) there are no functions in the hypothesis space that are not in B_N. Hence the concept space and the hypothesis space are equal and equation (23) is directly applicable.

h	5	10
N		
6	~ 230	~ 460
8	~ 900	~ 1800
10	~ 3560	~ 7200

Table 1: Values of $m_0 = h \times (2^N \cdot \log_e 2 + \log_e h)$

Proposition 4.8 Cardinality of B_N. *[AB92, p.6] The set of functions B_N has cardinality 2^{2^N}.*

Proof: The functions of B_N are defined on the domain D_N which has cardinality 2^N. Hence result since each bit may be mapped to 0 or 1 independently of all other mappings. □

Table 1 evaluates equation (23) for values of $N = 6, 8 \,\&\, 10$, and values of $h = 5 \,\&\, 10$ where we take $\delta = \epsilon$ and define $h = \frac{1}{\epsilon} = \frac{1}{\delta}$ so that the upper bound of equation (23) reduces to $m_0 = h(2^N \cdot \log_e 2 + \log_e h)$. The table shows how the value of m_0 increases with the size of hypothesis space and the required levels of confidence and accuracy. A sample size of about 230 is required to guarantee the learning of a function from B_6 with 80% accuracy and confidence (h = 5), while about 7200 training examples are needed to reach even 90% accuracy and confidence (h = 10) for a function from the larger space B_{10}. These numbers are clearly large compared to the number of distinct instances which could be stored in the case-base. (The maximum size of the case-base constructed by $CB1(\sigma)$ is $|D_N| = 2^N$.) It is noted that the sample size required by the Blumer bound is so high compared to the number of distinct instances because the PAC framework assumes that all members of the random sample are sampled independently, meaning that for a small, discrete space such as D_6, the number of repetitions will become large. Additionally, the sample complexity is a *worst-case* quantity, in that it refers to the size of training sample needed before the probability that the examples presented are pathologically unrepresentative of the target concept is acceptably small (within the confidence bound δ).

In giving *upper* bounds on sample complexity, equations (23) and (24) show that the size of training sample that can be processed before a consistent learning algorithm *necessarily* outputs a good hypothesis with high probability will increase with the cardinality and VC dimension of the hypothesis space of the learner. For the moment, it is assumed that the converse also holds, and that as the hypothesis space increases, the sample complexity of the learning algorithm also must increase. Equations (23) and (24) are taken as support for the intuition that, in general, the larger the hypothesis space, the more training examples the learner must see in order to discriminate between the available hypotheses, and choose a hypothesis that is accurate with high probability [Hau90, p.1103].

It is important to note that while arguments of this kind will be assumed in what follows, such a conclusion strictly depends on the specific properties of the learning algorithm using the hypothesis space. Also, in relating values for the sample complexity of a learning algorithm to the empirically derived learning curves for the average case described in previous sections, it is assumed that the 'worst case' quantity of sample complexity correlates with and is a useful predictor for the average case learning curve of a learning algorithm. While the results presented below agree in broad terms with these two assumptions, it is necessary to bear in mind that they remain assumptions and must be replaced by more careful analysis as progress is made.

Sample Complexity for Learning Monomial Functions $t \,\epsilon\, M_N$

Bearing in mind these caveats, the improved results of Figure 3 (learning M_6) compared with Figure 1 (learning B_6) are taken for the moment as an indication that for a restricted concept space such as M_N, only a fraction of the functions in B_N may be output as hypotheses. The smaller concept space M_N is learnt much more efficiently (with respect to sample size) than the space of all functions B_N. In general,

k	2^k	$\left(\begin{array}{c} N \\ k \end{array}\right)$	$\lvert M_{N,k}\rvert$
0	1	1	1
1	2	6	12
2	4	15	60
3	8	20	160
4	16	15	240
5	32	6	192
6	64	1	64

Table 2: Values of $|M_{N,k}| = 2^k \cdot {}^N C_k$ for $N = 6$

however, it cannot be assumed that the size $|H|$ of the hypothesis space of $CB1(\sigma)$ with respect to some concept space C necessarily varies with $|C|$. Consider the following results.

Proposition 4.9 Cardinality of M_N. *[AB92, p.12] The set of functions M_N has cardinality 3^N.*

Proof: Follows since, under some function from M_N, each bit of an N-vector from D_N may either be determined to be 1, determined to be 0 or undetermined. □

Proposition 4.10 Cardinality of $M_{N,k}$. *The set of functions $M_{N,k}$ has cardinality $2^k \cdot \left(\begin{array}{c} N \\ k \end{array}\right)$.*

$$|M_{N,k}| = 2^k \times \left(\begin{array}{c} N \\ k \end{array}\right)$$

Proof: We argue

$$|M_{N,k}| = \frac{2N \times 2(N-1) \times \ldots \times 2(N-k+1)}{k!}$$

since each literal in the monomial expression U may be either positive or negative, giving a choice of $2N$ possible literals on the first choice, $2(N-1)$ on the second choice, *etc.* The divisor allows for the many different permutations of the same set of literals. Hence $|M_{N,k}| = 2^k \cdot \left(\begin{array}{c} N \\ k \end{array}\right)$. □

Table 2 lists values of $|M_{N,k}|$ for $N = 6$ and values of k from 0 to 6. It is clear that $|M_{6,1}| < |M_{6,2}| < |M_{6,3}|$. A naive assumption that $|C|$ is going to correlate with $|H|$ would then predict that $M_{6,1}$ will be learnt the most easily and $M_{6,3}$ will be learnt the most slowly. Figure 4 clearly contradicts this, showing as it does that $CB1(\sigma_H)$ has the poorest learning behaviour for $M_{6,1}$ and the strongest for $M_{6,3}$. We must therefore consider more carefully what can be said of $H_C^{CB1(\sigma)}$, the hypothesis space of $CB1$ with respect to some given concept space C.

The observations of Figure 4 can be explained informally by considering that for a monomial target concept $t \in M_{N,k}$, there are 2^{N-k} descriptions in D_N which are positive instances of t. In the discussions of section 3, it was seen that the positive instances of a represented function are defined by the union of the 'regions of influence' of the positive exemplars in the case-base. (Consider for example equations (4) and (8)). Therefore, as the number of positive instances of the target concept increases, then so does the number of case-base representable functions. For case-bases derived from training samples in $M_{N,1}$ for example, equation (4) quantifies over 2^{N-1} possible positive exemplars, while for targets in $M_{N,N}$, the representable functions are those corresponding to the region of influence of a single positive exemplar. It might be expected then that the hypothesis space of $CB1(\sigma_H)$ with respect to the concept space $M_{N,k}$ will contain a greater variety of functions as k is decreased. Example 4.11 shows that this is indeed the case in that the effective hypothesis space $H_{M_{4,1}}^{CB1(\sigma_H)}$ contains functions which cannot be represented by case-bases extensible to target concepts in $H_{M_{4,2}}^{CB1(\sigma_H)}$ ($H_{M_{4,1}}^{CB1(\sigma_H)} \not\subseteq H_{M_{4,2}}^{CB1(\sigma_H)}$). However, the converse also holds. As k increases and the number of positive instances of a monomial target concept decreases, configurations of negative exemplars become possible which are also necessary to the representation of particular functions. Example 4.12 illustrates this with a function in $H_{M_{4,2}}^{CB1(\sigma_H)}$ which has no case-based representation with respect to targets in $M_{4,1}$. The arguments in these examples depend on the idea of the 'potentially relevant bits' of a case-base defined in Definition 4.9. Informally, the elements of $\Pi_{CB'}$ are the elements of the representation on which all positive exemplars in CB' agree. Therefore, any monomial representation U s.t. $CB' \subseteq h_U$ must be a subset of $\Pi_{CB'}$. Note that where $CB \subseteq h_U$ for some $h_U \,\epsilon\, M_{N,k}$, then necessarily $|\Pi_{CB}| \geq k$ since all positive exemplars must agree on the k bits constrained by U.

Definition 4.9 Potentially relevant bits with respect to a case-base CB *The potentially relevant bits of a representation with respect to a case-base* CB *are the bits on which the positive exemplars* $(d_{pos}, 1) \,\epsilon\, CB$ *are either all positive or all negative. We let* Π_{CB} *stand for the monomial expression defining the potentially relevant bits with respect to* CB*, i.e.* Π_{CB} *is the expression representing the most specific monomial function which has value 1 on all positive exemplars.*

$$\Pi_{CB} = \bigcap_{(d_{pos},1)\,\epsilon\, CB} U_{d_{pos}}$$

Example 4.11 *Consider the case-base* $CB = \{(1111,1), (1000,1), (0001,0), (0011,0), (0101,0), (0110,0), (0100,0), (0010,0)\}$*, and the function* $h_{\langle CB,\sigma_H\rangle}$ *represented by that case-base. Then:*

1. *The function* $h_{\langle CB,\sigma_H\rangle}$ *has boolean representation* $u_1.u_2.u_3.u_4 + u_1.\overline{u}_2.\overline{u}_3.\overline{u}_4$.
2. *The function* $h_{\langle CB,\sigma_H\rangle}$ *is a member of the hypothesis space of* $CB1(\sigma_H)$ *with respect to the set of monomials* $M_{4,1}$.
$$h_{\langle CB,\sigma_H\rangle} \,\epsilon\, H_{M_{4,1}}^{CB1(\sigma_H)} \tag{25}$$
3. but $h_{\langle CB,\sigma_H\rangle}$ *is* not *a member of the hypothesis space of* $CB1(\sigma_H)$ *with respect to* $M_{4,2}$.
$$h_{\langle CB,\sigma_H\rangle} \not\epsilon H_{M_{4,2}}^{CB1(\sigma_H)} \tag{26}$$

Proof: **1)** The boolean representation of $f = h_{\langle CB,\sigma_H\rangle}$ is (see p. 7):

$$\begin{aligned}&\Delta_{\sigma_H}(1111,0001).\Delta_{\sigma_H}(1111,0011).\Delta_{\sigma_H}(1111,0101).\Delta_{\sigma_H}(1111,0110).\Delta_{\sigma_H}(1111,0100).\Delta_{\sigma_H}(1111,0010)\\ &+ \Delta_{\sigma_H}(1000,0001).\Delta_{\sigma_H}(1000,0011).\Delta_{\sigma_H}(1000,0101).\Delta_{\sigma_H}(1000,0110).\Delta_{\sigma_H}(1000,0100).\Delta_{\sigma_H}(1000,0010)\end{aligned} \tag{27}$$

where, as in Example 3.4,

$$\Delta_{\sigma_H}(d_{pos}, d_{neg}) = \bigvee\{U \subseteq (U_{d_{pos}} \setminus U_{d_{neg}}) | \#U = \left\lfloor \frac{\#(U_{d_{pos}} \setminus U_{d_{neg}})}{2} \right\rfloor + 1\}$$

Hence:

$$\begin{aligned}&h_{\langle CB,\sigma_H\rangle}\\ &= (u_1u_2 + u_1u_3 + u_2u_3).(u_1u_2).(u_1u_3).(u_1u_4).(u_1u_3 + u_1u_4 + u_3u_4).(u_1u_2 + u_1u_4 + u_2u_4)\\ &\quad + (u_1\overline{u}_4).(u_1\overline{u}_3 + u_1\overline{u}_4 + \overline{u}_3\overline{u}_4).(u_1\overline{u}_2 + u_1\overline{u}_4 + \overline{u}_2\overline{u}_4).(u_1\overline{u}_2 + u_1\overline{u}_3 + \overline{u}_2\overline{u}_3).(u_1\overline{u}_2).(u_1\overline{u}_3)\\ &= (u_1u_2u_3u_4).(u_1u_2 + u_1u_3 + u_2u_3).(u_1u_3 + u_1u_4 + u_3u_4).(u_1u_2 + u_1u_4 + u_2u_4)\\ &\quad +(u_1\overline{u}_2\overline{u}_3\overline{u}_4).(u_1\overline{u}_3 + u_1\overline{u}_4 + \overline{u}_3\overline{u}_4)(u_1\overline{u}_2 + u_1\overline{u}_4 + \overline{u}_2\overline{u}_4).(u_1\overline{u}_2 + u_1\overline{u}_3 + \overline{u}_2\overline{u}_3).\\ &= (u_1u_2u_3u_4) + (u_1\overline{u}_2\overline{u}_3\overline{u}_4)\end{aligned}$$

2) Consider the function $h_{\{u_1\}} \epsilon M_{4,1}$. Now:

$$\begin{aligned}
h_{\{u_1\}}(1111) &= 1 \\
h_{\{u_1\}}(1000) &= 1 \\
h_{\{u_1\}}(0001) &= 0 \\
h_{\{u_1\}}(0011) &= 0 \\
h_{\{u_1\}}(0101) &= 0 \\
h_{\{u_1\}}(0110) &= 0 \\
h_{\{u_1\}}(0100) &= 0 \\
h_{\{u_1\}}(0010) &= 0
\end{aligned}$$

Hence $(d, n) \epsilon CB \rightarrow h_{\{u_1\}}(d) = n$, and $h_{\langle CB,\sigma_H\rangle} \epsilon H^{CB1(\sigma_H)}_{M_{4,1}}$.

3) Need to show

$$\forall t \epsilon M_{4,2} \cdot \forall CB' \subseteq t \cdot h_{\langle CB',\sigma_H\rangle} \neq u_1u_2u_3u_4 + u_1\overline{u}_2\overline{u}_3\overline{u}_4$$

Assume the negation, that there is some case-base CB extensible to a function in $M_{4,2}$ s.t. $h_{\langle CB,\sigma_H\rangle} = u_1u_2u_3u_4 + u_1\overline{u}_2\overline{u}_3\overline{u}_4$. Thus the point 1111 is a positive instance of $h_{\langle CB,\sigma_H\rangle}$, and this positive classification is due to a positive exemplar d_1^+ close to 1111 i.e. there is some $(d_1^+, 1) \epsilon CB$ s.t.

$$\forall (d^-, 0) \epsilon CB \cdot \sigma_H(1111, d_1^+) > \sigma_H(1111, d^-) \tag{28}$$

Assume $d_1^+ \neq 1111$. Take the least value x s.t. $(d_1^+)_x = 0$, and consider the instance d^* defined as follows:

$$\begin{aligned}
(d^*)_i &= 1 \quad i \neq x \\
(d^*)_x &= 0
\end{aligned}$$

Consider some bit i where $(d_1^+)_i = 1$; this cannot be the x-th bit and hence $(d^*)_i = 1$. Hence d_1^+ agrees with 1111 only if d^* agrees with d_1^+, and by contrapositive:

$$\forall 1 \leq i \leq N \cdot (d^*)_i \neq (d_1^+)_i \rightarrow (1111)_i \neq (d_1^+)_i \tag{29}$$

Hence by proposition 3.5, $h_{\langle CB,\sigma_H\rangle}(d^*) = 1$. But for no value of x will d^* be a positive instance of $u_1u_2u_3u_4 + u_1\overline{u}_2\overline{u}_3\overline{u}_4$, and so it must be concluded $d_1^+ = 1111$. By a similar argument there must be a second positive exemplar $(d_2^+, 1) \epsilon CB$ s.t. $d_2^+ = 1000$. Hence we have:

$$(1111, 1) \epsilon CB \tag{30}$$

$$(1000, 1) \epsilon CB \tag{31}$$

Thus $\Pi_{CB} \subseteq \{u_1\}$. But since CB is extensible to some target concept $t \epsilon M_{4,2}$, all positive exemplars in CB must agree on at least two bits of the representation; $|\Pi_{CB}| \geq 2$. Hence by contradiction it is inferred there is no such CB. □

Example 4.12 *Consider the case-base $CB = \{(1111, 1), (1100, 1), (1000, 0), (0100, 0)\}$, and the function $h_{\langle CB,\sigma_H\rangle}$ represented by that case-base. Then:*

1. *The function $h_{\langle CB,\sigma_H\rangle}$ has boolean representation $u_1.u_2 + u_3.u_4$.*
2. *The function $h_{\langle CB,\sigma_H\rangle}$ is a member of the hypothesis space of $CB1(\sigma_H)$ with respect to the set of monomials $M_{4,2}$.*

$$h_{\langle CB,\sigma_H\rangle} \epsilon H^{CB1(\sigma_H)}_{M_{4,2}} \tag{32}$$

3. but $h_{\langle CB,\sigma_H\rangle}$ is not *a member of the hypothesis space of* $CB1(\sigma_H)$ *with respect to* $M_{4,1}$.

$$h_{\langle CB,\sigma_H\rangle} \not\in H^{CB1(\sigma_H)}_{M_{4,1}} \tag{33}$$

Proof: **1)** The boolean representation of $f = h_{\langle CB,\sigma_H\rangle}$ is:

$$\begin{aligned}
&\Delta_{\sigma_H}(1111,1000).\Delta_{\sigma_H}(1111,0100) + \Delta_{\sigma_H}(1100,1000).\Delta_{\sigma_H}(1100,0100)\\
&= (u_2u_3 + u_2u_4 + u_3u_4)(u_1u_3 + u_1u_4 + u_3u_4) + (u_2)(u_1)\\
&= (u_1u_2u_3 + u_1u_2u_3u_4 + u_2u_3u_4 + u_1u_2u_3u_4 + u_1u_2u_4 + u_2u_3u_4 + u_1u_3u_4 + u_1u_3u_4 + u_3u_4) + u_1u_2\\
&= u_1u_2u_4 + u_1u_2u_3 + u_3u_4 + u_1u_2\\
&= u_1u_2 + u_3u_4
\end{aligned}$$

2) Consider the function $h_{\{u_1,u_2\}} \in M_{4,2}$. Now:

$$\begin{aligned}
h_{\{u_1,u_2\}}(1111) &= 1\\
h_{\{u_1,u_2\}}(1100) &= 1\\
h_{\{u_1,u_2\}}(1000) &= 0\\
h_{\{u_1,u_2\}}(0100) &= 0
\end{aligned}$$

Hence $(d,n) \in CB \rightarrow h_{\{u_1,u_2\}}(d) = n$, and $h_{\langle CB,\sigma_H\rangle} \in H^{CB1(\sigma_H)}_{M_{4,2}}$.

3) Need to show

$$\forall t \in M_{4,1} \cdot \forall CB' \subseteq t \cdot h_{\langle CB',\sigma_H\rangle} \neq u_1u_2 + u_3u_4$$

Assume the negation, that there is some case-base CB extensible to a function in $M_{4,1}$ s.t. $h_{\langle CB,\sigma_H\rangle} = u_1u_2 + u_3u_4$. For ease of notation, define the monomial expressions $U_1 = \{u_1, u_2\}$ and $U_2 = \{u_3, u_4\}$. Hence:

$$\forall d \in D_N \cdot (\exists (d_{pos}, 1) \in CB \cdot \forall (d_{neg}, 0) \in CB \cdot \sigma_H(d, d_{pos}) > \sigma_H(d, d_{neg})) \leftrightarrow (h_{U_1}(d) = 1 \vee h_{U_2}(d) = 1) \tag{34}$$

Since σ_H is a definite similarity measure, $(d,n) \in CB \rightarrow h_{\langle CB,\sigma_H\rangle}(d) = n$ and:

$$\forall (d_{pos}, 1) \in CB \cdot h_{U_1}(d_{pos}) = 1 \vee h_{U_2}(d_{pos}) = 1 \tag{35}$$

Additionally, consider the point 1100 which is a positive instance of $h_{\langle CB,\sigma_H\rangle}$. Clearly the exemplars maximally similar to 1100 are positive ones; let any such exemplar be denoted d_1^+. Assume that d_1^+ disagrees with 1100 on the first or the second bit of the description i.e. $(d_1^+)_1 = 0 \vee (d_1^+)_2 = 0$. Assume the former ($(d_1^+)_1 = 0$) and consider the description 0100; clearly $(1100)_i = (d_1^+)_i$ only if $(0100)_i = (d_1^+)_i$ and by proposition 3.5, $h_{\langle CB,\sigma_H\rangle}(0100) = 1$, which we know to be false. A similar contradiction is derived with respect to the classification of 1000 if we assume $(d_1^+)_2 = 0$. Hence:

$$\exists (d_{pos}, 1) \in CB \cdot h_{U_1}(d_{pos}) = 1 \tag{36}$$

By a similar consideration of the description 0011, we conclude:

$$\exists (d_{pos}, 1) \in CB \cdot h_{U_2}(d_{pos}) = 1 \tag{37}$$

Also CB is extensible to some target concept $t \in M_{4,1}$. Hence all positive exemplars in CB must agree on at least one bit of the representation ($\Pi_{CB} \neq \emptyset$). In addition, from equations (36) & (37), it is clear that any bits on which the positive exemplars all agree will have the value 1 in those exemplars. Hence:

$$\exists 1 \leq i \leq 4 \cdot \forall (d_{pos}, 1) \in CB \cdot (d_{pos})_i = 1 \tag{38}$$

$$\forall 1 \leq i \leq 4 \cdot u_i \in \Pi_{CB} \rightarrow \forall (d_{pos}, 1) \in CB \cdot (d_{pos})_i = 1 \tag{39}$$

$$\forall 1 \leq i \leq 4 \cdot u_i \not\in \Pi_{CB} \rightarrow \exists (d_{pos}, 1) \in CB \cdot (d_{pos})_i = 0 \tag{40}$$

As a final preliminary, note that the monomial target concept from which CB contains instances must be defined by a subset of Π_{CB}; hence with respect to the negative exemplars in the case-base:

$$\exists \Pi \subseteq \Pi_{CB} \cdot \# \Pi = 1 \wedge \forall (d_{neg}, 0) \in CB \cdot h_{\Pi}(d_{neg}) = 0 \tag{41}$$

Case 1). Assume that the set of potentially relevant literals Π_{CB} is such that for some $l \in \{1, 2\}$, exactly one of the literals in U_l also is in Π_{CB}. :

$$\exists l \in \{1, 2\} \cdot (\exists u_i \in U_l \cdot u_i \in \Pi_{CB}) \wedge (\exists u_i \in U_l \cdot u_i \not\in \Pi_{CB}) \tag{42}$$

Let l be any such value, and let l' be its complement i.e. the value $l' \in \{1, 2\} \cdot l' \neq l$. Let i be the index of the literal in U_l s.t. $u_i \not\in \Pi_{CB}$ and let i' be the index of the other literal in U_l. By equations (39) & (40), there must be some $(d_{pos}, 1) \in CB$ such that $(d_{pos})_{i'} = 1$ and $(d_{pos})_i = 0$. In addition, from equation (35), $h_{U_{l'}}(d_{pos}) = 1$. Finally, let i'' denote a bit of the representation s.t. $u_{i''} \in U_{l'}$ and $i'' \neq j$, where j is the least index s.t. $\Pi = \{u_j\}$ satisfies equation (41).

Consider the description d_1 which differs from d_{pos} only on bit i''. Clearly this is a negative instance of $h_{\langle CB, \sigma_H \rangle}$ since $h_{U_l}(d_1) = 0$ and $h_{U_{l'}}(d_1) = 0$. From equation (34), there must be some d_{neg} at least as similar to d_1 as the d_{pos} adjacent to d_1. Hence:

$$\exists (d_{neg}, 0) \in CB \cdot \sigma_H(d_1, d_{neg}) \geq \frac{3}{4} \tag{43}$$

Now j is a potentially relevant bit, so $(d_{pos})_j = 1$. d_1 differs from d_{pos} only on bit i'' which is distinct from j, so $(d_1)_j = 1$. However, from equation (41), $(d_{neg})_j = 0$, and since $\sigma_H(d_1, d_{neg}) \geq \frac{3}{4}$, d_1 and d_{neg} must differ precisely on this bit. Consider a final description d_2 adjacent to d_{neg} but is a positive instance of $h_{\langle CB, \sigma_H \rangle}$. Note there is exactly one such description. d_{pos} by definition had 3 bits set, and exactly two of these have become set to 0 in the definition of d_{neg}, leaving a single bit of d_{neg} with value 1. By the process of the proof, if $u_j \in U_l$, then this bit will be left in $U_{l'}$, while if $u_j \in U_{l'}$ then this bit will be in U_l. Hence there is one bit that can be set so that d_2 is a positive instance of either h_{U_1} or h_{U_2}. This bit will not be j however, since u_j will be in the half of the description where both bits of d_{neg} are zero. Hence $(d_2)_j = (d_{neg})_j = 0$. But j is a potentially relevant bit, so by equation (39), $(d_2, 1)$ cannot be a positive exemplar in CB, and the adjacency of d_{neg} requires that $h_{\langle CB, \sigma_H \rangle}(d_2) = 0$, contradicting the assumption that d_2 is a positive instance of the represented function.

Hence there is no case-based representation of $u_1 u_2 + u_3 u_4$ in $H_{M_{4,1}}^{CB1(\sigma_H)}$ satisfying the assumption (42).

Case 2) Assume instead of equation (42) that Π_{CB} matches exactly one of U_1 and U_2:

$$\exists l, l' \in \{1, 2\} \cdot (\Pi_{CB} \cap U_l = U_l) \wedge (\Pi_{CB} \cap U_{l'} = \emptyset) \tag{44}$$

Let l & l' be such values. Let d_1 be the positive instance of $h_{\langle CB, \sigma_H \rangle}$ such that $(d_1)_{i'} = 1$ for $u_{i'} \in U_{l'}$ and $(d_1)_i = 0$ for $u_i \in U_l$. By equation (39), all positive exemplars $(d_{pos}, 1) \in CB$ will have $(d_{pos})_i = 1$ for all $u_i \in \Pi_{CB}$; since $\Pi_{CB} = U_l$, we have therefore:

$$\forall (d_{pos}, 1) \in CB \cdot \sigma_H(d_1, d_{pos}) \leq \frac{1}{2} \tag{45}$$

Since d_1 is a positive instance of $h_{\langle CB,\sigma_H\rangle}$, then it follows in addition:

$$\forall (d_{neg},0)\,\epsilon\, CB \cdot \sigma_H(d_1,d_{neg}) < \frac{1}{2} \tag{46}$$

Note that this allows no more than two distinct negative exemplars to be present in CB. That is, a description d_{neg} may be a negative exemplar only if the following hold, as demonstrated below:

$$\#\{u_i\,\epsilon\, U_{l'} | (d_{neg})_i = 1\} = 0 \tag{47}$$

$$\#\{u_i\,\epsilon\, U_l | (d_{neg})_i = 1\} = 1 \tag{48}$$

Assume that equation (47) does not hold. Either $h_{U_{l'}}(d_{neg}) = 1$, which would make d_{neg} a positive instance of $h_{\langle CB,\sigma_H\rangle}$, or there is exactly one $u_i\,\epsilon\, U_{l'}$ s.t. $(d_{neg})_i = 1$. In this case, either there is no more than one literal $u_i\,\epsilon\, U_l$ s.t. $(d_{neg})_i = 1$, giving $\sigma_H(d_1,d_{neg}) \geq \frac{1}{2}$, or $h_{U_l}(d_{neg}) = 1$. Thus $(d_{neg},0)\,\epsilon\, CB$ requires (47). Assume instead equation (48) does not hold. Either $h_{U_l}(d_{neg}) = 1$ or d_{neg} agrees with d_1 on the two bits of U_l, giving $\sigma_H(d_1,d_{neg}) \geq \frac{1}{2}$. Thus d_{neg} cannot be a negative exemplar in this case either.

Take some description d_{neg} satisfying equations (47) & (48), and assume it is not a negative exemplar. Consider a description d_2 adjacent to d_{neg}, differing from d_{neg} only on some bit i s.t. $u_{i'}\,\epsilon\, U_{l'}$. Now equations (36), (37), (39) & (44) require that there is a positive exemplar $(1111,1)\,\epsilon\, CB$, and $\sigma_H(1111,d_2) = \frac{1}{2}$. The only negative exemplar in CB will differ from d_2 on both bits in U_l and additionally the bit i in $U_{l'}$ on which d_2 and d_{neg} differ, giving a similarity to d_2 of $\frac{1}{4}$. Hence d_2 will be a positive instance of $h_{\langle CB,\sigma_H\rangle}$. But d_2 has one bit in U_l set and one bit in $U_{l'}$; hence $h_{U_l}(d_2) = 0$ and $h_{U_{l'}}(d_2) = 0$, requiring that d_2 is a negative instance. Hence $(d_{neg},0)$ must be a negative exemplar in CB for either d_{neg} satisfying equations (47) & (48). But since $\Pi_{CB} = U_l$, there must be at least one bit j in U_l which is zero in all negative exemplars $(d_{neg},0)$ (equation (41)). Since this allows only one exemplar satisfying (47) & (48) to be in the case-base, it is concluded there is no case-based representation for $u_1u_2 + u_3u_4$ in $H_{M_{4,1}}^{CB1(\sigma_H)}$ satisfying equation (44).

Case 3) Assume that neither equation (42) nor (44) holds. Given the negation of (42), either $|U_l \cap \Pi_{CB}| = 2$ or $|U_l \cap \Pi_{CB}| = 0$ for $l\,\epsilon\,\{1,2\}$. From (38) and the negation of (44), we infer that all bits must be potentially relevant. Thus the case analysis can be completed by considering the case $|\Pi_{CB}| = 4$, requiring a unique positive exemplar $(d_{pos},1)\,\epsilon\, CB$. Therefore the argument above can be repeated with respect to the positive instances 1100 and 0011 giving contradictory constraints on any negative exemplars in the case-base. Hence there can be no such negative exemplars and no case-based representation for $u_1u_2 + u_3u_4$ in $H_{M_{4,1}}^{CB1(\sigma_H)}$ s.t. $|\Pi_{CB}| = 4$. □

While a direct enumeration of $H_{M_{N,k}}^{CB1(\sigma_H)}$ for small values of N shows that the number and variety of representable hypotheses does indeed increase as k decreases for fixed N, Examples 4.11 and 4.12 show that a simple characterisation such as $H_{M_{N,k}}^{CB1(\sigma_H)} \supseteq H_{M_{N,k'}}^{CB1(\sigma_H)}$ for $k \leq k'$ does not hold. Indeed, a direct characterisation of $d_{VC}(H_{M_{N,k}}^{CB1(\sigma_H)})$ or $|H_{M_{N,k}}^{CB1(\sigma_H)}|$ as a function on N and k has not yet been achieved, and an 'explanation' in these terms for the differentiated learning curves in Figure 4 is not yet available. Proposition 4.17 below, however, gives one positive result about these 'effective hypothesis spaces'. Although this is very much a partial characterisation of the spaces, defining as it does only a subset of the representable hypotheses, it is reported because of conclusions that it suggests about the optimality of the sample complexity of $CB1(\sigma_H)$. These conclusions will be developed below once the result has been established.

Specifically, Proposition 4.17 shows that given training samples only for monomial target concepts of a fixed size $t\,\epsilon\, M_{N,k}$, then there are training samples which will cause CB1 to output on that training sample any one of the monomial functions $h\,\epsilon\, M_N$; the result shows that for all N and k, $M_N \subseteq H_{M_{N,k}}^{CB1(\sigma_H)}$. The result is shown straightforwardly by induction on the size of the representation N. A number of preliminaries are required therefore to deal with relationships between functions and representations as the order of the example space increases.

Firstly, we consider the functions in B_{N+1} whose monomial expressions differ from the representation of a specific monomial function $h_U \,\epsilon\, M_N$ only on the 'extending' bit of the representation $N+1$.

Definition 4.10 Extrapolations of monomial function. *The extrapolations of a monomial function $h_U^N \,\epsilon\, M_N$ are the functions $h_{U'}^{N+1} \,\epsilon\, M_{N+1}$ such that $U' \,\epsilon\, \{U, U \cup \{u_{N+1}\}, U \cup \{\overline{u_{N+1}}\}\}$.*

$$h_{U'}^{N+1} \,\epsilon\, \mathit{extr}_N(h_U^N) \leftrightarrow (U' = U \vee U' = U \cup \{u_{N+1}\} \vee U' = U \cup \{\overline{u_{N+1}}\})$$

Proposition 4.13 *The union of the extrapolations of the functions $f \,\epsilon\, M_N$ is equal to the class of functions M_{N+1}.*

$$\forall N \geq 0 \cdot \bigcup_{f \,\epsilon\, M_N} \mathit{extr}_N(f) = M_{N+1}$$

Proof: Clearly $\bigcup_{f \,\epsilon\, M_N} \mathrm{extr}_N(f) \subseteq M_{N+1}$ since extr_N contains only functions from M_{N+1}. Also $M_{N+1} \subseteq \bigcup_{f \,\epsilon\, M_N} \mathrm{extr}_N(f)$ since for every function $h_{U'}^{N+1}$ there is a function h_U^N where $U = U' \setminus \{u_{N+1}, \overline{u_{N+1}}\}$ whose extrapolation contains $h_{U'}^{N+1}$. □

In a similar way, we must consider the 'projections' of a description $d \,\epsilon\, D_N$ as the representation is extended:

Definition 4.11 Projections of a description. *The projections of a description are constructed by extending the description by a single new bit.*

$$\forall d \,\epsilon\, D_N, d' \,\epsilon\, D_{N+1} \cdot d' \,\epsilon\, \mathit{proj}_N(d) \leftrightarrow \forall 1 \leq i \leq N \cdot (d)_i = (d')i$$

Definition 4.12 Projections of a case-base. *The projections of a case-base are constructed by adding a new bit, set to one specified value, to the description of each exemplar in the case-base.*

$$P_i^N(CB) = \{(d', n) | (d, n) \,\epsilon\, CB \wedge d' \,\epsilon\, \mathit{proj}_N(d) \wedge (d')_{N+1} = i\}$$

A single boolean expression such as $u_1u_2 + u_2u_3$ represents different functions if it is interpreted as a function on different sized example spaces D_N and $D_{N'}$. Similarly, there are case-based representations of these functions which are also very similar to one another. The definitions of 'projections' allow simple results along these lines to be expressed below. That is, consider a boolean expression Φ that, when interpreted as a function on the N-dimensional example space D_N, represents a function $f^N \,\epsilon\, B_N$ which also has case-based representation $\langle CB, \sigma_H \rangle$. Proposition 4.15 below states that either projection of CB (Definition 4.12) will give a case-based representation of the function defined on D_{N+1} by the same boolean expression Φ. For example, given $\langle CB, \sigma_H \rangle$ as a representation for the function in B_3 represented by $u_1u_2 + u_1u_3$, then both $\langle P_0^3(CB), \sigma_H \rangle$ and $\langle P_1^3(CB), \sigma_H \rangle$ will represent the function in B_4 represented by $u_1u_2 + u_1u_3$. The result is given in two parts with Proposition 4.14 serving as a lemma for the proof of Proposition 4.15.

Proposition 4.14 *Given a function $f^{N+1} \,\epsilon\, B_{N+1}$ defined on D_{N+1} and a second function $h_\Phi^N \,\epsilon\, B_N$ defined on D_N and represented by the boolean form Φ, it is concluded that $f^{N+1} = h_\Phi^{N+1}$, i.e. f^{N+1} is the function on D_{N+1} represented by the same boolean expression Φ, if it can be shown that for any description $d \,\epsilon\, D_N$, $h_\Phi^N(d)$ will return the same value as $f^{N+1}(d')$, where d' is either of the projections of d in D_{N+1}.*

$$\begin{aligned}&\forall N \geq 1 \cdot \forall f^{N+1} \,\epsilon\, B_{N+1}, h_\Phi^N \,\epsilon\, B_N \cdot \\ &\quad (\forall d \,\epsilon\, D_N, d' \,\epsilon\, D_{N+1} \cdot d' \,\epsilon\, \mathit{proj}_N(d) \rightarrow (h_\Phi^N(d) = 1 \leftrightarrow f^{N+1}(d') = 1)) \rightarrow f^{N+1} = h_\Phi^{N+1}\end{aligned}$$

Proof: For some pair of functions f^{N+1} & h_Φ^N, assume the LHS of the implication:

$$\forall d \,\epsilon\, D_N, d' \,\epsilon\, D_{N+1} \cdot d' \,\epsilon\, \mathrm{proj}_N(d) \rightarrow (h_\Phi^N(d) = 1 \leftrightarrow f^{N+1}(d') = 1) \tag{49}$$

It will be shown that $f^{N+1} \neq h_\Phi^{N+1}$ requires a contradiction, hence the result. Assume $f^{N+1} \neq h_\Phi^{N+1}$:

$$\exists d' \,\epsilon\, D_{N+1} \cdot f^{N+1}(d') \neq h_\Phi^{N+1}(d') \tag{50}$$

Since Φ depends only on the first N bits of the representation, it is clear that $h_\Phi^{N+1}(d') = 1 \leftrightarrow h_\Phi^N(d') = 1$, where d is the unique $d \,\epsilon\, D_N$ s.t. $d' \,\epsilon\, \mathrm{proj}_N(d)$ and hence d & d' agree on the first N bits of the representation. Hence from (50) we have a $d \,\epsilon\, D_N$ & $d' \,\epsilon\, \mathrm{proj}_N(d)$ contradicting (49) and (49) entails $f^{N+1} = h_\Phi^{N+1}$. □

Proposition 4.15 *Consider a function $f^N \,\epsilon\, B_N$ which is represented by the boolean form Φ and by the case-based representation $\langle CB, \sigma_H\rangle$. Then the function $f^{N+1} \,\epsilon\, B_{N+1}$ represented by the same boolean form Φ will be represented by both $\langle P_0^N(CB), \sigma_H\rangle$ and $\langle P_1^N(CB), \sigma_H\rangle$.*

Proof: It must be shown that for any function $h^N_{\langle CB,\sigma_H\rangle}(d)$ defined on D_N the functions represented by the projections of CB will classify the projections of d positively iff $h^N_{\langle CB,\sigma_H\rangle}(d) = 1$:

$$\begin{aligned} &\forall N \geq 1 \cdot \forall i \,\epsilon\, \{0,1\} \cdot \forall h^N_{\langle CB,\sigma_H\rangle} \,\epsilon\, B_N \cdot \\ &\qquad \forall d \,\epsilon\, D_N, d' \,\epsilon\, \mathrm{proj}_N(d) \cdot (h^N_{\langle CB,\sigma_H\rangle}(d) = 1 \leftrightarrow h^{N+1}_{\langle P_i^N(CB),\sigma_H\rangle}(d') = 1) \end{aligned} \tag{51}$$

Then, given Proposition 4.14, it will follow immediately that the functions $h^N_{\langle CB,\sigma_H\rangle}$ and $h^{N+1}_{\langle P_i^N(CB),\sigma_H\rangle}$ may be represented by the same boolean form Φ.

Assume there is some $d \,\epsilon\, D_N$ such that $h_{\langle CB,\sigma_H\rangle}(d) = 1$, and let d' be a projection of d in D_{N+1}. There must be a positive exemplar in CB satisfying equation (4). For any $d_1, d_2, d_3 \,\epsilon\, D_N$ where $\sigma_H(d_1, d_2) > \sigma_H(d_1, d_3)$, consider the projections of d_1 in D_{N+1}, $d_1' \,\epsilon\, \mathrm{proj}_N(d_1)$. Consider also projections of d_2 & d_3, $d_2' \,\epsilon\, \mathrm{proj}_N(d_2), d_3' \,\epsilon\, \mathrm{proj}_N(d_3)$, such that $(d_2')_{N+1} = (d_3')_{N+1}$. Let $\gamma_{i,j}$ stand for the number of bits which d_i and d_j agree on; similarly, let $\gamma_{i',j'}$ stand for the number of bits agreed on by d_i' and d_j'. Since the extending bit $(d_1')_{N+1}$ will either agree or disagree with the bit extending d_2 and d_3, we have $\gamma_{1',2'} - \gamma_{1,2} = \gamma_{1',3'} - \gamma_{1,3} = \delta$, where $\delta \,\epsilon\, \{0,1\}$. Therefore we also have $\sigma_H(d_1', d_2') > \sigma_H(d_1', d_3')$, and, letting $d' = d_1'$, any d_{pos} from the projection of the case-base $= d_2'$ and any $d_{neg} = d_3'$:

$$\forall i \,\epsilon\, \{0,1\} \cdot \exists (d_{pos}, 1) \,\epsilon\, P_i^N(CB) \cdot \forall (d_{neg}, 0) \,\epsilon\, P_i^N(CB) \cdot \sigma_H(d', d_{pos}) > \sigma_H(d', d_{neg}) \tag{52}$$

and $h^{N+1}_{\langle P_i^N(CB),\sigma_H\rangle}(d') = 1$, $i \,\epsilon\, \{0,1\}$. Similarly $h^N_{\langle CB,\sigma_H\rangle}(d) = 0 \rightarrow h^{N+1}_{\langle P_i^N(CB),\sigma_H\rangle}(d') = 0$. Hence (51), and as noted above, the result follows immediately from Proposition 4.14. □

By way of final preliminary, we note the following, which will allow the induction on N to proceed independently of the number of relevant bits k.

Proposition 4.16 *For a given case base CB containing exactly one positive exemplar, if there is a function $f \,\epsilon\, M_{N,k}$ s.t. $CB \subseteq f$, then for any larger k' s.t. $k \leq k' \leq N$, there is some $f' \,\epsilon\, M_{N,k'}$ so that also $CB \subseteq f'$.*

$$\begin{aligned} &\forall N \geq 1 \cdot \forall 1 \leq k \leq N \cdot \forall CB \,\epsilon\, \mathcal{P}\ (D_N \times \{0,1\}) \cdot \\ &\qquad (\#\{d_{pos} :\ D_N | (d_{pos}, 1) \,\epsilon\, CB\} = 1 \rightarrow \\ &\qquad\qquad \forall f \,\epsilon\, M_{N,k} \cdot CB \subseteq f \rightarrow \forall k \leq k' \leq N \cdot \exists f' \,\epsilon\, M_{N,k'} \cdot CB \subseteq f') \end{aligned}$$

Proof: Assume some N & k, a target concept $f \epsilon M_{N,k}$ and CB s.t. CB contains a single positive exemplar and in addition $CB \subseteq f$. Consider the monomial representation U s.t. $h_U^N = f$. Consider also any monomial function $h_{U'} \epsilon M_{N,k'}$ s.t. $U \subseteq U'$ and additionally $h_{U'}(d^+) = 1$ where d^+ is the one description in D_N s.t. $(d^+, 1) \epsilon CB$. Clearly there will be $\left(\begin{array}{c} N-k \\ k'-k \end{array} \right)$ such functions, since extra literals must be chosen referring to $k' - k$ out of the $N - k$ unconstrained bits. Hence trivially $\forall (d_{pos}, 1) \epsilon CB \cdot h_{U'}(d_{pos}) = 1$. Further, any negative exemplars $(d_{neg}, 0) \epsilon CB$ will be negative instances of h_U and hence negative instances of the more specific $h_{U'}$. Therefore $\forall (d_{neg}, 0) \epsilon CB \cdot h_{U'}(d_{neg}) = 0$, and $CB \subseteq h_{U'}$. □

Hence:

Proposition 4.17 *The effective hypothesis space* $H_{M_{N,k}}^{CB1(\sigma_H)}$ *of the case-based learning algorithm* $CB1(\sigma_H)$*, defined with respect to the 'unweighted feature count' similarity measure* σ_H *and the set of k-literal monomial functions* $M_{N,k}$*, contains the set of all monomial functions* M_N *defined on* D_N*.*

$$\forall N \geq 1 \cdot \forall 1 \leq k \leq N \cdot M_N \subseteq H_{M_{N,k}}^{CB1(\sigma_H)}$$

Proof: By induction on N. Proposition 4.1 shows that the required result is equivalent to requiring that for each $f \epsilon M_N$, there is a 'target concept' $t \epsilon M_{N,k}$ for any value $1 \leq k \leq N$, and some case base $CB \subseteq t$, such that $h_{\langle CB, \sigma_H \rangle}^N = f$. Therefore, it will be sufficient to show $\forall N \geq 1 \cdot H(N)$, defining H as below. Introducing the extra restriction that case bases contain a single positive exemplar will allow reference to proposition 4.16 in subsequent argument:

$$H(N) \hat{=} \forall f \epsilon M_N \cdot \forall 1 \leq k \leq N \cdot \exists t \epsilon M_{N,k} \cdot \exists CB \subseteq t \cdot p^+(CB) \wedge h_{\langle CB, \sigma_H \rangle}^N = f$$

where $p^+(CB) \hat{=} \#\{d_{pos} : D_N | (d_{pos}, 1) \epsilon CB\} = 1$.

Base Case $H(1)$. $M_1 = \{\{\}, \{u_1\}, \{\overline{u_1}\}\}$. $h_{\langle \{(1,1)\}, \sigma_H \rangle} = h_{\{\}}$, $h_{\langle \{(1,1),(0,0)\}, \sigma_H \rangle} = h_{\{u_1\}}$ and $h_{\langle \{(1,0),(0,1)\}, \sigma_H \rangle} = h_{\{\overline{u_1}\}}$. Hence $H(1)$.

Inductive Step $H(p) \rightarrow H(p+1)$. We make the inductive hypothesis $H(p)$:

$$\forall f \epsilon M_p \cdot \forall 1 \leq k \leq p \cdot \exists t \epsilon M_{p,k} \cdot \exists CB \subseteq t \cdot p^+(CB) \wedge h_{\langle CB, \sigma_H \rangle}^p = f \tag{53}$$

Proposition 4.13 indicates that it will be sufficient to infer from equation (53) that for any monomial function $f \epsilon M_p$ each extrapolation of f is a member of the hypothesis space with respect to $M_{p+1,k}$ for values $1 \leq k \leq p+1$. Proposition 4.16 in turn shows that it will be sufficient to derive from the inductive hypothesis that for each $f' \epsilon \mathrm{extr}_p(f)$ there is a $t \epsilon M_{p+1,1}$ and a case-base $CB \subseteq t$ containing just one positive exemplar which represents f', which will entail the results for all other values of k.

Hence it will be shown equation (53) entails that for each $h_U^p \epsilon M_p$ there are functions t_1, t_2 and t_3 and case bases CB_1, CB_2 and CB_3 satisfying:

$$\forall h_U^p \epsilon M_p \cdot \exists t_1 \epsilon M_{p+1,1} \cdot \exists CB_1 \subseteq t_1 \cdot p^+(CB_1) \wedge h_{\langle CB_1, \sigma_H \rangle}^{p+1} = h_U^{p+1} \tag{54}$$

$$\forall h_U^p \epsilon M_p \cdot \exists t_2 \epsilon M_{p+1,1} \cdot \exists CB_2 \subseteq t_2 \cdot p^+(CB_2) \wedge h_{\langle CB_2, \sigma_H \rangle}^{p+1} = h_{U \cup \{u_{p+1}\}}^{p+1} \tag{55}$$

$$\forall h_U^p \epsilon M_p \cdot \exists t_3 \epsilon M_{p+1,1} \cdot \exists CB_3 \subseteq t_3 \cdot p^+(CB_3) \wedge h_{\langle CB_3, \sigma_H \rangle}^{p+1} = h_{U \cup \{\overline{u_{p+1}}\}}^{p+1} \tag{56}$$

For any function $h_U^p \epsilon M_p$, equation (53) asserts there must be some case base CB s.t. there is some $h_T^p \epsilon M_{p,1}$ where $CB \subseteq h_T^p$ and $h_{\langle CB, \sigma_H \rangle}^p = h_U^p$. It will be shown that there are case-bases defined in terms of CB and T which will satisfy each of equations (54) to (56):

a) Case-based representation of h_U^{p+1}*.*

Given the inductive hypothesis, from Proposition 4.15 it can be inferred immediately that $P_0^p(CB)$ and $P_1^p(CB)$ are case-based representation of h_U^{p+1}:

$$h^{p+1}_{\langle P_0^p(CB),\sigma_H\rangle} = h^{p+1}_{\langle P_1^p(CB),\sigma_H\rangle} = h_U^{p+1} \tag{57}$$

Note also that $P_i^p(CB)$ will contain only a single positive exemplar. To establish equation (54), it remains only to show $P_i^p(CB) \subseteq h_T^{p+1}$. For any $(d',n)\,\epsilon\, P_i^p(CB)$, there is a unique d such that $d'\,\epsilon\, \text{proj}_p(d)$ and $(d,n)\,\epsilon\, CB$ (definition 4.12). Since $CB \subseteq h_T^p$, $(d,n)\,\epsilon\, CB \rightarrow h_T^p(d) = n$. Since d & d' agree on their first p bits and also $h_T^p\,\epsilon\, M_{p,k}$ so that T refers only to the first p bits of representation, $h_T^p(d) = 1 \leftrightarrow h_T^{p+1}(d') = 1$. Hence also $h_T^{p+1}(d') = n$ and therefore $(d',n)\,\epsilon\, P_i^p(CB) \rightarrow h_T^{p+1}(d') = n$. Hence the following result, concluding (54):

$$P_i^p(CB) \subseteq h_T^{p+1} \tag{58}$$

b) Case-based representation of $h^{p+1}_{U\cup\{u_{p+1}\}}$. It will be shown that the case base $P_1^p(CB) \cup \{(d_{new},0)\}$ is a case-based representation of $h^{p+1}_{U\cup\{u_{p+1}\}}$, where d_{new} is defined as follows:

$$\begin{aligned}
(d_{new})_x &= |1 - (d_{pos})_x| \\
(d_{new})_i &= (d_{pos})_i \;\; \textit{where} \;\; 1 \leq i \leq p \;\wedge\; i \neq x \\
(d_{new})_{p+1} &= 0
\end{aligned}$$

d_{pos} is the description of the unique positive exemplar in $P_1^p(CB)$, inherited from CB, and x is the value $1 \leq x \leq N$ s.t. $u_x\,\epsilon\, T \vee \overline{u_x}\,\epsilon\, T$, T being the representation of the target function $h_T^p\,\epsilon\, M_{p,1}$.

By equation (4), we have $h^{p+1}_{\langle P_1^N(CB)\cup\{(d_{new},0)\},\sigma_H\rangle} = f^{p+1}$, where:

$$f^{p+1}(d) = \begin{cases} 1 & \textit{if } h^{p+1}_{\langle P_1^p(CB),\sigma_H\rangle}(d) = 1 \wedge \sigma_H(d,d_{pos}) > \sigma_H(d,d_{new}) \\ 0 & \textit{otherwise} \end{cases} \tag{59}$$

From the definition of d_{new}, we have $\sigma_H(d,d_{pos}) > \sigma_H(d,d_{new})$ iff d agrees with d_{pos} on a strict majority of the bits $\{u_x, u_{p+1}\}$; note $(d_{pos})_{p+1} = 1$ since $(d_{pos},1)\,\epsilon\, P_1^p(CB)$, while $(d_{new})_{p+1} = 0$ by definition. (All other bits are irrelevant to the comparison since they are common to both d_{pos} and d_{new}). Hence:

$$\sigma_H(d,d_{pos}) > \sigma_H(d,d_{new}) \leftrightarrow ((d)_x = (d_{pos})_x \wedge (d)_{p+1} = 1) \tag{60}$$

Substituting (57) and (60) in (59):

$$h^{p+1}_{\langle P_1^p(CB)\cup\{(d_{new},0)\},\sigma_H\rangle} = f^{p+1} = h^{p+1}_{U\cup\{u_{p+1}\}} \tag{61}$$

since $h_U^{p+1}(d) = 1$ implies that $(d)_x$ must have the same value as $(d_{pos})_x$.

Clearly, the new case-base still contains a single positive exemplar; to satisfy equation (55), it must only be shown $P_i^p(CB) \cup \{(d_{new},0)\} \subseteq h_T^{p+1}$. From (58), we have $P_i^p(CB) \subseteq h_T^{p+1}$. Note also $h_T^{p+1}(d_{new}) = 0$ since by definition, d_{new} will fail to satisfy T. Hence $(d_{new},0)\,\epsilon\, h_T^{p+1}$ and $P_1^p(CB) \cup \{(d_{new},0)\} \subseteq h_T^{p+1}$.

c) Case-based representation of $h^{p+1}_{U\cup\{\overline{u_{p+1}}\}}$. Equally, the case base $P_0^p(CB) \cup \{(d'_{new},0)\}$, where $d'_{new} = d_{new}$ except $(d'_{new})_{p+1} = 1$ (d_{new} defined as above), is an equivalent representation to $U \cup \{\overline{u_{p+1}}\}$. Hence (56). □

As an immediate corollary of Proposition 4.17, we have:

Corollary 4.18 Lower bound on VC Dimension of $H_{M_{N,k}}^{CB1(\sigma_H)}$. *The VC Dimension of $H_{M_{N,k}}^{CB1(\sigma_H)}$, the effective hypothesis space of $CB1(\sigma_H)$ with respect to the concept space $M_{N,k}$, is at least of the order of N.*

$$d_{VC}(H_{M_{N,k}}^{CB1(\sigma_H)}) = \Omega(N)$$

Proof: $H_{M_{N,k}}^{CB1(\sigma_H)}$ contains M_N (proposition 4.17). Therefore any sample shattered by M_N will be shattered by $H_{M_{N,k}}^{CB1(\sigma_H)}$, and the VC dimension of $H_{M_{N,k}}^{CB1(\sigma_H)}$ will be at least that of M_N, which is $\mathcal{O}(N)$ [AB92, p.76] [Hau88, p.193]. □

In contrast to corollary 4.18, note the following results:

Proposition 4.19 Upper bound on VC Dimension of $M_{N,k}$. *The VC Dimension of $M_{N,k}$ is no greater than* $1 + \log_2 \left(\begin{array}{c} N \\ k \end{array} \right)$.

Proof: Let $\overline{x}$ be a sample of size v, which orders the set of examples X and is shattered by $M_{N,k}$. Consider that there are 2^{v-1} subsets of X which contain a particular $x_i \,\epsilon\, X$, and also that there are exactly $\left(\begin{array}{c} N \\ k \end{array} \right)$ functions $f \,\epsilon\, M_{N,k}$ that classify x_i positively. Since each subset of X must be labelled by a distinct member of $M_{N,k}$, we have $2^{v-1} \leq \left(\begin{array}{c} N \\ k \end{array} \right)$, and hence $v \leq 1 + \log_2 \left(\begin{array}{c} N \\ k \end{array} \right)$ □

Proposition 4.20 Lower bound on VC Dimension of $M_{N,k}$. *The VC Dimension of the set of functions $M_{N,k}$ is at least $1 + \lfloor \log_2(N - k + 1) \rfloor$.*

Proof: Proof will be by demonstrating the construction of a sample $\overline{x} = \langle x_1, \ldots, x_m \rangle$ of size $m = 1 + \lfloor \log_2(N - k + 1) \rfloor$, and showing that it is shattered by $M_{N,k}$. Let $l = \lfloor \log_2(N - k + 1) \rfloor$, so that 2^l is the largest power of 2 no greater than $(N - k + 1)$. Consider the powerset $\mathcal{P}\ S$ where S is the set of tokens $\{t_1, \ldots, t_{l+1}\}$; an enumeration over the powerset is assumed so that the elements may be referred to as $S_1, \ldots, S_{2^{l+1}}$. The enumeration is also required to have the property that for any pair of disjoint subsets whose union is S, their indices must have the sum $2^{l+1} + 1$:

$$\forall S_j, S_{j'} \subseteq S \cdot S_j \cap S_{j'} = \emptyset \wedge S_j \cup S_{j'} = S \rightarrow j + j' - 1 = 2^{l+1} \tag{62}$$

It is asserted this condition can be satisfied without loss of generality. To define $\overline{x}$, set the jth bit, $1 \leq j \leq N$ of example x_i, $1 \leq i \leq l + 1$, as follows:

$$\begin{array}{rlll} (x_i)_j = & 1 & \mathit{if}\, t_i \,\epsilon\, S_j & 1 \leq j \leq 2^l \\ (x_i)_j = & 0 & \mathit{if}\, t_i \not\epsilon S_j & 1 \leq j \leq 2^l \\ (x_i)_j = & 1 & 2^l < j \leq N & \end{array}$$

Since $2^l \leq (N - k + 1)$, this definition reserves at least $k - 1$ bits of the representation and sets them to 1; the remaining bits are defined s.t. bit j of the representation has value 1 for any example x_i if and only if the corresponding token t_i belongs to S_j, some specific subset of S.

The proof must now show how there exists a total, injective function f mapping between $\mathcal{P}\ S$ and the set of functions $M_{N,k}$ such that

$$\forall S_j \,\epsilon\, \mathcal{P}\ S : \ f(S_j)(x_i) = 1 \leftrightarrow t_i \,\epsilon\, S_j; \tag{63}$$

this would demonstrate the shattering of $\overline{x}$ by $M_{N,k}$. $|S| = l + 1$, so values of j in the range $1 \leq j \leq 2^{l+1}$ must be accounted for. The subranges $1 \leq j \leq 2^l$ and $2^l + 1 \leq l \leq 2^{l+1}$ are treated as separate cases:

$1 \leq j \leq 2^l$. For some S_j therefore, note $(x_i)_j = 1$ iff $t_i \epsilon S_j$. Therefore equation (63) is satisfied by the k-literal monomial function m such that $m(x) = 1$ iff the value 1 is assigned to $(x)_j$ and the $k-1$ bits $(x)_{j'}$, $2^l < j' \leq 2^l + k - 1$.

$2^l + 1 \leq j \leq 2^{l+1}$. For these S_j, note that there is a unique subset $S_{j'}$ s.t. $S_{j'} = S \setminus S_j$ and by equation (62) $j' = 2^{l+1} - j + 1$. Hence $1 \leq j' \leq 2^l$ and $(x_i)_{j'} = 1$ iff $t_i \epsilon S_{j'}$. Since in addition it is clear that $t_i \epsilon S_{j'} \leftrightarrow t_i \not\epsilon S_j$, we have $(x_i)_{j'} = 0 \leftrightarrow t_i \epsilon S_j$. Therefore equation (63) is satisfied by $m \epsilon M_{N,k}$ such that $m(x) = 1$ iff $(x)_j$ has the value 0 and the $k-1$ bits $(x)_{j'}$, $2^l < j' \leq 2^l + k - 1$, all have value 1. □

Proposition 4.21 VC Dimension of $M_{N,1}$. *The VC Dimension of the set of functions* $M_{N,1}$ *is* $1 + \lfloor \log_2(N) \rfloor$.

Proof: From proposition 4.19 VCDim($M_{N,1}$) $\leq 1 + \log_2(N)$, while from proposition 4.20, VCDim($M_{N,1}$) $\geq 1 + \lfloor \log_2(N) \rfloor$. □

While an exact result is possible for the case of $k = 1$, the bounds of propositions 4.19 and 4.20 are not tight for other values of k (except where $k = N$ and $d_{VC}(M_{N,N}) = 1$). However, for all $1 \leq k < N$, we have upper and lower bounds on $d_{VC}(M_{N,k})$ which increase in $\log N$, even though the relationship to k is not clear. (We have however explored the maximal shattered samples with respect to $M_{N,2}$ and $M_{N,3}$ by search of the sample space; the data generated is consistent with the hypothesis that the actual VC dimension fits the upper bound to within a constant, increasing linearly in k' where $k' = \min(k, N-k)$ as well as in $\log N$. It has not been possible to establish this analytically, however.)

Hence, while the VC dimension of the hypothesis space of $CB1(\sigma_H)$ with respect to the set of functions $M_{N,k}$ is at least of the order of N (Corollary 4.18), the VC dimension of $M_{N,k}$ itself is $\mathcal{O}(\log N)$ (Proposition 4.19 and 4.20). Equations (23) & (24) (and the discussion of p. 17 ff) lead us to believe that this qualitative difference indicates that $CB1(\sigma_H)$ is a less than optimal learning algorithm (with respect to sample complexity) for the space $M_{N,k}$. That is, as N increases, we expect that the number of examples $CB1(\sigma_H)$ needs in order to reach an accurate hypothesis will rapidly outgrow the number needed by a learning algorithm whose hypothesis space represents exactly the functions contained in $M_{N,k}$.

4.4 Conclusions: Learning Behaviour of $CB1(\sigma_H)$

Considering the result of Corollary 4.6, it is clear that $CB1(\sigma_H)$ is a general purpose learning algorithm with a rich hypothesis language. Specifically, for any fixed *definite* similarity measure (Definition 4.3) such as σ_H, corollaries 4.6 and 4.4 indicate that there is a case-based representation $\langle CB, \sigma_H \rangle$ for any $\{0,1\}$-valued total function on D_N. In addition, proposition 4.1 shows that for any such representation, there is a training sample that causes $CB1(\sigma_H)$ to output that representation. Thus corollary 4.5 states formally that $CB1(\sigma_H)$ is a general purpose learning algorithm in the sense of the PAC learning framework, 'probably' able to generate an arbitrarily good approximation to any classification of the example space provided that enough examples are available to the learner.

However, in considering the sample complexity of $CB1(\sigma_H)$, it has been shown here (Proposition 4.17) that the hypothesis space of $CB1(\sigma_H)$ with respect to the concept space $M_{N,k}$ includes not only $M_{N,k}$ but also *all* monomial functions M_N. In addition to that formal result, direct enumeration establishes the presence of functions such as $u_1 + u_2.u_3$ and $u_1.u_2 + u_1.u_3 + u_2.u_3$ in $H^{CB1(\sigma_H)}_{M_{3,1}}$ and that $|H^{CB1(\sigma_H)}_{M_{N,k}}|$ does indeed increase as k decreases. Unfortunately, the partial characterisation of $H^{CB1(\sigma_H)}_{M_{N,k}}$ of Proposition 4.17 does not yet allow us to account directly for the differentiation of the learning curves in Figure 4 in terms of differences in the hypothesis space. However, arguments related to equations (23) and (24) lead us to believe in addition that, for all values of k, the presence of these spurious hypotheses will make $CB1(\sigma_H)$ a relatively inefficient learning algorithm for $M_{N,k}$ (with respect to sample complexity) compared to a consistent learning algorithm which can represent *only* the functions $M_{N,k}$.

We suggest that this is a natural corollary of the generality of $CB1(\sigma_H)$; this seems a clear example of the concept of inductive bias [Hau88] [Sch94]. 'Bias' refers to any prior information or knowledge that might be encoded in a learning algorithm that defines a preference for choosing a hypothesis from the many that might be available to account for the training data. "If a bias is strong and correct, then the concept-learning task is relatively easy because the concept learner will be guided to the selection of the target concept" [Utg86, p.114], while "[if no inductive bias] is supplied for comparing competing hypotheses [consistent with the available exemplars], than all possible classifications of the unseen instances are equally possible and no inductive method can do better on average than random guessing" [Hau88, p.178]. Altering the 'bias' of a learning algorithm allows a trade-off between generality and sample complexity to be managed; more bias will lead to greater accuracy for the same size of training sample, but the increased bias will only be correct for a small number of possible target concepts. On the other hand, as bias is weakened, the learner will be able to output hypotheses approximating a wider range of target concepts, but will, on average, require more examples before converging to a *good* approximation.

It seems clear then, that the choice of σ_H as similarity measure instantiates $CB1(\sigma)$ as a *low bias* learning algorithm that will successfully learn a wide range of target concepts at the cost of generally high sample complexity. This is sometimes seen as *characteristic* of case-based learning, a tendency reflected in descriptions of case-based reasoning as a paradigm suitable for *knowledge-poor* domains e.g. [CCK93]. This stance seems incorrect; in contrast, Wess and Globig have already pointed out and ably demonstrated that "the [similarity] measure (respectively the way to modify the measure) is the bias of case-based reasoning" [WG94, p.90]. That is, with some prior knowledge of the concept space to be learnt, the similarity measure can be manipulated so that the hypotheses output by the case-based learner are more likely to be close to the possible target concepts. Such strategies demonstrably improve efficiency with respect to sample size [GW94] [WG94], although performance will obviously be degraded outside the chosen concept space. These issues will be explored further in the following section.

5 Case-Based Learning with Variable Similarity Measures

$CB1(\sigma)$, considered in the previous section, is a simple case-based learning algorithm that learns with a fixed similarity measure and adjusts its hypothesis only by adding cases to the case-base. Globig & Wess assert, however, that "in a case-based learner, two processes - reducing the size of the learnable concepts (hypothesis space) and increasing the size of the case-base - should be performed"[WG94, p.88]. In this section, therefore, we wish to investigate how the hypothesis space can be manipulated directly by the choice of similarity measure. This will be done by defining $CB2$, a variant of $CB1(\sigma)$ which we assume to be equipped with a certain, ideal, similarity measure. As before, our main concern is to establish the sample complexity of $CB2$, and the method used will be the same, via an exploration of the 'effective hypothesis space' of $CB2$ with respect to various classes of target concept. Thus §5.1 reports our investigation of the class of functions having a case-based representation with respect to the ideally weighted similarity measure $\sigma_{\overline{w}_t}$, §5.2 proves the consistency of $CB2$, §5.3 reports on the empirically observed average case learning behaviour of CB2 and finally §5.4 presents results on the sample complexity of $CB2$. The main results to be presented are:

- In §5.1, we first consider whether a case-base of exemplars for a monomial target function $t \in M_N$ actually gives a representation of t in the sense defined within this section. A precise characterisation of such case-bases is given in Proposition 5.1. We are then able to give a straightforward method (Definition 5.4 and Proposition 5.5) for establishing whether a particular function $f \in B_N$ may be output by $CB2$ on a training sample for a monomial target concept. In the remainder of this first subsection we then develop two necessary conditions for membership of the hypothesis space of $CB2$ with respect to monomial target concepts $t \in M_N$; together these conditions give a reasonable bound on this space of functions (Proposition 5.15).

- In §5.4, in contrast to Examples 4.11 & 4.12 we are able to show that $H^{CB2}_{M_{N,k}} \subseteq H^{CB2}_{M_{N,k'}}$ for $k \leq k'$ (Proposition 5.20), indicating that the upper bound on the sample complexity of $CB2$ given by equations (23) and (24) must increase with k. Then by considering two different concepts of a 'covering net' [AA91, p 554] [KV94, pp 57-58], we are able to present various upper bounds on the sample complexity of $CB2$ (Corollary 5.23 & Corollary 5.26).

An algorithm suitable for learning M_N which both constructs a case-base and manipulates the similarity measure is presented by Globig & Wess [WG94, Fig 4], given below as definition 5.1. The definition refers to the weighted feature sum $\sigma_{\overline{w}}$ of definition 3.2.

Definition 5.1 VS-CBR Learning Algorithm for Functions in M_N *c.f. [WG94, Fig 4]*

```
define the functions $f_i : \{0,1\} \to \{0,1\}$, s.t. $f_i(n) = 1$, $1 \leq i \leq N, n \,\epsilon\, \{0,1\}$
set $CB = \emptyset$
for i = 1 to m do
    if $b_i = 1$ then
        if $\neg\exists d \,\epsilon\, D_N \cdot (d,1) \,\epsilon\, CB$ then set $CB = CB \cup \{(x_i,1)\}$
        for j = 1 to $N$ do
            set $f_j(1-(x_i)_j) = 0$
    else
        set $CB = CB \cup \{(x_i,0)\}$
for i = 1 to N do
    if $f_i(0) = 1 \vee f_i(1) = 1$ then
        set $w_i = 1$
    else
        set $w_i = 0$
set $VS\text{-}CBR(\overline{s}) = h_{\langle CB, \sigma_{\overline{w}} \rangle}$
```

where $\overline{s} = \langle (x_i, b_i) \rangle_{i=1}^{m}$ is a training sample from $(D_N \times \{0,1\})^m$.

This apparently elaborate algorithm can be understood in relation to the 'standard learning algorithm for monomials' [Val84a]. A single positive exemplar is kept as a 'prototype' of the monomial concept. In the same fashion as in the standard monomial algorithm, the other positive exemplars are used to determine whether or not a specific bit is necessary to the definition of the concept. This information is recorded in the vector of functions f_i. After processing any sample $\overline{s}$, $f_i(n) = 1$ only if no positive exemplar d_{pos} has been processed such that $(d_{pos})_i = n$. This means that in converting the f vector to the weight vector $\overline{w}$, VS-CBR can infer that a bit of the representation is irrelevant to the definition of the concept whenever both possible values have been seen in positive exemplars and hence $f_i(0) = 0 \wedge f_i(1) = 0$. In this case, the corresponding weight w_i is set to 0 so that bit i is ignored in the feature sum; otherwise, it is set to 1.

Rather than attempting to analyse this relatively complex algorithm, we define a related but more straightforward strategy for learning using the case-based representation. The similarity measure in the hypothesis chosen by $CB2$, defined below, weights bits as relevant or irrelevant according to whether they are determined by the target concept. This is clearly the ideal being approximated by VS-CBR, in that the ideally weighted similarity measure $\sigma_{\overline{w}_t}$ will eventually be inferred by VS-CBR given enough positive instances of the target concept.

Definition 5.2 CB2 Learning Algorithm for Case-Based Classifiers

```
set $CB = \emptyset$
for i = 1 to m do
    set $CB = CB \cup \{(x_i, b_i)\}$
set $CB2(\overline{s}) = h_{\langle CB, \sigma_{\overline{w}_t} \rangle}$
```

where $\overline{s} = \langle (x_i, b_i) \rangle_{i=1}^{m}$ is a training sample from $(D_N \times \{0,1\})^m$ and the weight vector $\overline{w}_t$ is defined

$$(w_t)_i = \begin{cases} 1 & \text{if } u_i \,\epsilon\, U_t \vee \overline{u_i} \,\epsilon\, U_t \\ 0 & \text{otherwise} \end{cases} \tag{64}$$

and where U_t is the monomial expression s.t. $t = h_{U_t}$.

Obviously, while VS-CBR is executable in that it constructs a weight vector to mask out irrelevant bits only so far as these can be inferred from the training sample, $CB2$ 'cheats' by defining such a weight vector with perfect knowledge of the target concept. This clearly is not an algorithmic learning strategy, but has the advantage that it greatly simplifies the representation of hypotheses. The results presented in this section are, as a direct consequence, somewhat more substantial than those in §4. This definition is also useful because it separates the purely case-based part of learning, the collection of cases in the case-base, from the process of inferring the relevant bits for the similarity measure. The latter is, as shown above, merely an instance of an otherwise very well known algorithm. This separation then demonstrates in the limit the maximum possible contribution to case-based learning that might be made by this approach of changing weights in the similarity measure. This contribution is assessed in the conclusions of §5.4 and §6.

5.1 Case-based Representation using the Weighted Feature Count $\sigma_{\overline{w}_t}$

Definition 5.3 *A function $f \,\epsilon\, B_N$ is* $CB2$**-representable** *with respect to some target concept $t \,\epsilon\, M_N$ if there is a case-base CB extensible to t ($CB \subseteq t$) s.t. $\langle CB, \sigma_{\overline{w}_t} \rangle$ is a case-based representation for f ($h_{\langle CB, \sigma_{\overline{w}_t} \rangle} = f$).*

This section will consider the problem of characterising the functions that are $CB2$-representable with respect to some target concept t. Note that $CB2$ (definition 5.2) is defined specifically for learning monomial functions, so only target concepts $t \,\epsilon\, M_N$ are considered. For convenience, H_t will be used to denote the set of functions that are $CB2$-representable with respect to a target concept t.

$$H_t = \{h_{\langle CB, \sigma_{\overline{w}_t} \rangle} | CB \subseteq t\} \tag{65}$$

Introducing zero weights into the similarity measure as in definitions 5.1 and 5.2 has the result that descriptions $d, d' \,\epsilon\, D_N$ which differ only on zero-weighted bits of the representation are equivalent for the purposes of classification. As in [WG94], let $\approx_\sigma$ define an equivalence relation s.t. $d_1 \approx_\sigma d_2 \leftrightarrow \sigma(d_1, d_2) = 1$. As Wess and Globig [WG94, p.86] state, the same classification must be shared by all descriptions in each equivalence class in the partition $\langle D_N \backslash \approx_\sigma \rangle$ for any function defined by the similarity function σ, and in addition the similarity of two descriptions in such an equivalence class to any $d \,\epsilon\, D_N$ will always be the same:

$$\forall d \,\epsilon\, D_N, D \,\epsilon\, \langle D_N \backslash \approx_\sigma \rangle, d_1, d_2 \,\epsilon\, D \cdot \sigma(d_1, d) = \sigma(d_2, d) \tag{66}$$

In the special case of a monomial target function $t \,\epsilon\, M_N$, then also *all* positive instances of t will lie in the same class of the partition $\langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle$ defined by the 'ideal' similarity measure $\sigma_{\overline{w}_t}$. This has the result that the case-based semantics of equation (4) are simplified. In the general case of equation (4), the case-based semantics nests the comparison of similarities within an existential quantification over the positive exemplars in the case-base; the interpretation of the case-base in the worst case must therefore consider every positive exemplar. Given a case-base CB extensible to some target concept $t \,\epsilon\, M_N$ however, the interpretation of the representation $\langle CB, \sigma_{\overline{w}_t} \rangle$ need only consider similarities with respect to a single positive instance of the target concept. This is a result of equation (66), since the quantification over negative exemplars $(\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_{\overline{w}_t}(d, d_{pos}) > \sigma_{\overline{w}_t}(d, d_{neg}))$ will be satisfied with respect to some positive exemplar d_{pos} if and only if it is satisfied with respect to all positive instances of the target concept. This allows a number of results to be proved about representations $\langle CB, \sigma_{\overline{w}_t} \rangle$ using the ideally weighted similarity measure $\sigma_{\overline{w}_t}$ which do not hold in the general case of representations containing the unweighted feature count σ_H.

Counting the Case-Based Representations of $t \in M_N$

Firstly, it is possible to define conditions necessary and sufficient to establish that a case-base CB is a representation, with respect to the weighted similarity measure $\sigma_{\overline{w}_t}$, of the target concept t. The number of exact representations for a given target concept may then be counted. The relevant results are expressed below as propositions 5.1 & 5.4.

Proposition 5.1 *For any monomial target concept $t \in M_{N,k}$, then a case-base $CB \subseteq t$ which is extensible to t is a representation of the target concept t with respect to the weighted similarity measure $\sigma_{\overline{w}_t}$ ($h_{\langle CB,\sigma_{\overline{w}_t}\rangle} = t$) if and only if the case-base CB contains:*

1. *At least one positive exemplar, and,*

2. *Negative exemplars s.t. for every relevant bit j s.t. $\{u_j, \overline{u}_j\} \cap U_t \neq \emptyset$, then given the description $d^+_{j\mapsto 0}$ defined as follows in terms of any positive instance of the target concept d^+ s.t. $t(d^+) = 1$:*

$$\begin{aligned}(d^+_{j\mapsto 0})_i &= (d^+)_i \quad i \neq j \\ (d^+_{j\mapsto 0})_j &= 1 - (d^+)_j\end{aligned}$$

then there is a negative exemplar $(d^-, 0) \in CB$ s.t. $\sigma_{\overline{w}_t}(d^+_{j\mapsto 0}, d^-) \geq \frac{k-1}{k}$.

$$\forall 1 \leq k \leq N \cdot \forall t \in M_{N,k} \cdot \forall CB \subseteq t \cdot h_{\langle CB,\sigma_{\overline{w}_t}\rangle} = t \leftrightarrow$$
$$[(\exists (d_{pos}, 1) \in CB) \wedge (\forall 1 \leq j \leq N \cdot \{u_j, \overline{u}_j\} \cap U_t \neq \emptyset \to \exists (d_{neg}, 0) \in CB \cdot \sigma_{\overline{w}_t}(d^+_{j\mapsto 0}, d_{neg}) \geq \frac{k-1}{k})]$$

Proof: **1) Only if:** Assume that $h_{\langle CB,\sigma_{\overline{w}_t}\rangle} = t$. Since there is at least one positive instance d^+ of any monomial function $t \in M_N$, then $h_{\langle CB,\sigma_{\overline{w}_t}\rangle}$ has value 1 on this description and hence there must have been at least one positive exemplar in CB. Consider instead some bit j relevant to the definition of the target concept ($\{u_j, \overline{u}_j\} \cap U_t \neq \emptyset$) and the 'nearly positive' instance $d^+_{j\mapsto 0}$ defined in terms of a positive instance d^+ as above. Now $t(d^+_{j\mapsto 0}) = 0$, but $\sigma_{\overline{w}_t}(d^+_{j\mapsto 0}, d_{pos}) = \frac{k-1}{k}$ for any positive exemplar $(d_{pos}, 1)$ in the case-base. But since $h_{\langle CB,\overline{w}_t\rangle} = t$, then $h_{\langle CB,\sigma_{\overline{w}_t}\rangle}(d^+_{j\mapsto 0}) = 0$, and hence $\forall (d_{pos}, 1) \in CB \cdot \exists (d_{neg}, 0) \in CB \cdot \sigma_{\overline{w}_t}(d^+_{j\mapsto 0}, d_{pos}) \leq \sigma_{\overline{w}_t}(d^+_{j\mapsto 0}, d_{neg})$, and hence $\exists (d_{neg}, 0) \in CB \cdot \sigma_{\overline{w}_t}(d^+_{j\mapsto 0}, d_{neg}) \geq \frac{k-1}{k}$.

2) If: Assume the RHS of the expression, so that for some case-base $CB \subseteq t$:

$$\exists d_{pos} \in D_N \cdot (d_{pos}, 1) \in CB \tag{67}$$
$$\forall 1 \leq j \leq N \cdot \{u_j, \overline{u}_j\} \cap U_t \neq \emptyset \to \exists d_{neg} \in D_N \cdot (d_{neg}, 0) \in CB \wedge \sigma_{\overline{w}_t}(d^+_{j\mapsto 0}, d_{neg}) \geq \frac{k-1}{k} \tag{68}$$

It must be shown that for any such case-base CB, $h_{\langle CB,\sigma_{\overline{w}_t}\rangle} = t$. We wish to show initially $h_{\langle CB,\sigma_{\overline{w}_t}\rangle}(d) = 1 \to t(d) = 1$. For a proof by contradiction, take a description $d \in D_N$ which is a positive instance of $h_{\langle CB,\sigma_{\overline{w}_t}\rangle}$ and assume $t(d) = 0$. Let d^+ denote the description of any positive instance of t. Now if $t(d) = 0$ then d must disagree with d^+ on δ 'relevant' bits of the representation, s.t. $1 \leq \delta \leq k$. Hence $\sigma_{\overline{w}_t}(d, d^+) = \frac{k-\delta}{k}$, and by equation (4) then for any negative exemplar in the case-base:

$$\forall (d_{neg}, 0) \in CB \cdot \sigma_{\overline{w}_t}(d, d_{neg}) < \frac{k-\delta}{k} \tag{69}$$

Now, consider the description $d^+_{j\mapsto 0}$ where j is the least index s.t. j is a relevant bit on which d & d^+ disagree. Now d and $d^+_{j\mapsto 0}$ must disagree on precisely $\delta - 1$ relevant bits, giving $\sigma_{\overline{w}_t}(d, d^+_{j\mapsto 0}) = \frac{k-\delta+1}{k}$. But, from equation (68), there is a negative exemplar $(d_{neg}, 0) \in CB$ s.t. $\sigma_{\overline{w}_t}(d^+_{j\mapsto 0}, d_{neg}) \geq \frac{k-1}{k}$. Hence by the 'transitivity' of $\sigma_{\overline{w}}$ (Proposition 3.1), we have $\sigma_{\overline{w}_t}(d, d_{neg}) \geq \frac{k-\delta}{k}$, contradicting equation (69). Hence $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(d) = 1 \rightarrow t(d) = 1$.

Assume instead that $d \,\epsilon\, D_N$ is a positive instance of t. Clearly $\sigma_{\overline{w}_t}(d, d_{pos}) = 1$, for the positive exemplar $(d_{pos}, 1) \,\epsilon\, CB$ required by equation (67). Additionally, since any negative instances of t will disagree with d on at least one relevant bit, then $\sigma_{\overline{w}_t}(d, d_{neg}) < 1$ for any negative exemplar $(d_{neg}, 0) \,\epsilon\, CB$. Hence $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(d) = 1$. Thus we have $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(d) = 1 \leftrightarrow t(d) = 1$. □

Proposition 5.1 allows the number of representations of a particular target concept $t \,\epsilon\, M_{N,k}$ to be counted straightforwardly and expressed as a function of N and k. For some such function, we define the quantity N_t to be the number of case-bases $CB \subseteq t$ s.t. $\langle CB, \sigma_{\overline{w}_t} \rangle$ is a case-based representation for the target concept $t \,\epsilon\, M_{N,k}$.

$$N_t \mathrel{\hat{=}} \#\{CB \subseteq t | h_{\langle CB, \sigma_{\overline{w}_t} \rangle} = t\} \tag{70}$$

Consider the partition $\langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle$ defined on the example space as above. Let D_1 be the set of those equivalence classes from $\langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle$ s.t. all descriptions in $D' \,\epsilon\, D_1$ differ from the positive instances of t on exactly one relevant bit of the representation, and D_2 be those equivalence classes from $\langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle$ whose members disagree with the positive instances on two relevant bits.

$$D_1 \mathrel{\hat{=}} \{D' \,\epsilon\, \langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle | \forall d \,\epsilon\, D', d_{pos} \,\epsilon\, D_N \cdot t(d_{pos}) = 1 \rightarrow \sigma_{\overline{w}_t}(d, d_{pos}) = \frac{k-1}{k}\} \tag{71}$$

$$D_2 \mathrel{\hat{=}} \{D' \,\epsilon\, \langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle | \forall d \,\epsilon\, D', d_{pos} \,\epsilon\, D_N \cdot t(d_{pos}) = 1 \rightarrow \sigma_{\overline{w}_t}(d, d_{pos}) = \frac{k-2}{k}\} \tag{72}$$

Now by proposition 5.1, whether or not $h_{\langle CB, \sigma_{\overline{w}_t} \rangle} = t$ depends only on the presence of a positive exemplar, and some specific negative exemplars, all of which differ from the positive exemplars on no more than two of the relevant bits of the representation i.e. $(d_{neg}, 0) \,\epsilon\, CB$ s.t. $\sigma_{\overline{w}_t}(d_{neg}, d_{pos}) \geq \frac{k-2}{k}$. Specifically, for every $D' \,\epsilon\, D_1$, either D' is non-empty (i.e. $\forall (d, n) \,\epsilon\, CB \cdot d \not\epsilon D'$), or there is one of the equivalence classes $D'' \,\epsilon\, D_2$ adjacent to D' ($\forall d' \,\epsilon\, D', d'' \,\epsilon\, D'' \cdot \sigma_{\overline{w}_t}(d', d'') = \frac{k-1}{k}$) which is non-empty.

Given a specific subset of D_1 s.t. i of the equivalence classes in $D' \,\epsilon\, D_1$ (out of the total of k) are 'empty', i.e. *not* covered by some exemplar in the case-base ($\forall (d, n) \,\epsilon\, CB \cdot d \not\epsilon D'$), then let $Z_{k,i}$ be the number of subsets of D_2 that may be chosen such that if precisely those $D' \,\epsilon\, D_2$ are non-empty (containing some exemplar of the case-base), then the function represented by $\langle CB, \sigma_{\overline{w}_t} \rangle$ will be precisely t (provided that the case-base also contains some positive exemplar).

$$Z_{k,i} \mathrel{\hat{=}} \#\{D \subseteq D_2 | (\forall D'' \,\epsilon\, D_2 \cdot D'' \,\epsilon\, D \leftrightarrow \exists d \,\epsilon\, D'' \cdot (d, 0) \,\epsilon\, CB) \rightarrow h_{\langle CB, \sigma_{\overline{w}_t} \rangle} = t\} \tag{73}$$

Lemma 5.2 Pascal's Triangle. *e.g. [AB92, p.78]*

$$\forall a, b \geq 0 \cdot \left(\begin{array}{c} a+1 \\ b+1 \end{array} \right) = \left(\begin{array}{c} a \\ b+1 \end{array} \right) + \left(\begin{array}{c} a \\ b \end{array} \right)$$

Proposition 5.3 *The value of $Z_{k,i}$ is defined for all $0 \leq i \leq k$ as follows:*

$$Z_{k,i} = \sum_{x=0}^{i} (-1)^x . \left(\begin{array}{c} i \\ x \end{array} \right) .2^{\left(\begin{array}{c} k-x \\ 2 \end{array} \right)}$$

Proof: by induction on i. **BaseCase** $i = 0$. If all $D' \,\epsilon\, D_1$ are hit by some exemplar in the case-base, then any of the $2^{\left(\begin{array}{c} k \\ 2 \end{array} \right)}$ subsets of D_2 will be consistent with a representation of t. As defined above $Z_{k,0} = 1 \times 1 \times 2^{\left(\begin{array}{c} k \\ 2 \end{array} \right)}$. **Inductive Step.** Assume that $Z_{k,p} = \sum_{x=0}^{p} (-1)^x . \left(\begin{array}{c} p \\ x \end{array} \right) .2^{\left(\begin{array}{c} k-x \\ 2 \end{array} \right)}$ for some value $0 \leq p < k$. It must be shown that $Z_{k,p+1} = \sum_{x=0}^{p+1} (-1)^x . \left(\begin{array}{c} p+1 \\ x \end{array} \right) .2^{\left(\begin{array}{c} k-x \\ 2 \end{array} \right)}$.

Take a set Δ_{p+1} of $p+1$ classes from D_1 (e.g. those classes $D' \epsilon D_1$ which differ from the positive instances of t on the first $p+1$ relevant bits of the representation). Similarly let Δ_p be the set of classes from D_1 whose descriptions disagree with positive instances of t on the first p bits of the representation. Assume that all classes $D' \epsilon \Delta_{p+1}$ are empty.

Note there are $Z_{k,p}$ subsets of D_2 which, when non-empty, are sufficient for the representation of t given that all classes $D' \epsilon \Delta_p$ are empty. However, not all these $Z_{k,p}$ subsets will contain a class $D'' \epsilon D_2$ whose descriptions are adjacent to those of the class in $\Delta_{p+1} \setminus \Delta_p$. Thus $Z_{k,p+1} = Z_{k,p} - Y$, where Y is the number of subsets of D_2 satisfying the property $Y_p(\Delta)$, i.e. Δ is a subset of D_2 which covers the empty classes in Δ_p, but which do not contain any $D'' \epsilon D_2$ whose descriptions are adjacent to those of the class in $\Delta_{p+1} \setminus \Delta_p$.

Consider the set $D_2' = \{D'' \epsilon D_2 | \forall D' \epsilon D_1 \cdot \{D'\} = \Delta_{p+1} \setminus \Delta_p \to \forall d' \epsilon D', d'' \epsilon D'' \cdot \sigma_{\overline{w}_t}(d', d'') < \frac{k-1}{k}\}$, i.e. the set of elements of D_2 whose descriptions are not adjacent to the extra class in $\Delta_{p+1} \setminus \Delta_p$. By definition, all Δ satisfying $Y_p(\Delta)$ are also in D_2'. Thus we need only count the number of elements in D_2' which cover Δ_p. This is clearly the same problem as counting $Z_{k-1,p}$. Hence

$$Z_{k,p+1} = Z_{k,p} - Z_{k-1,p} \qquad (74)$$

Thus from the inductive hypothesis:

$$\begin{aligned}
Z_{k,p+1} &= Z_{k,p} - Z_{k-1,p} \\
&= \sum_{x=0}^{p}(-1)^x \binom{p}{x} 2^{\binom{k-x}{2}} - \sum_{x=0}^{p}(-1)^x \binom{p}{x} 2^{\binom{k-x-1}{2}} \\
&= \sum_{x=0}^{p}(-1)^x \binom{p}{x} 2^{\binom{k-x}{2}} - \sum_{x'=1}^{p+1}(-1)^{x'-1} \binom{p}{x'-1} 2^{\binom{k-x'}{2}} \\
&= 2^{\binom{k}{2}} + \sum_{x=1}^{p} 2^{\binom{k-x}{2}} \left[(-1)^x \binom{p}{x} - (-1)^{x-1} \binom{p}{x-1}\right] - (-1)^p 2^{\binom{k-p-1}{2}} \\
&= 2^{\binom{k}{2}} + \sum_{x=1}^{p}(-1)^x \left[\binom{p}{x} + \binom{p}{x-1}\right] 2^{\binom{k-x}{2}} + (-1)^{p+1} 2^{\binom{k-p-1}{2}}
\end{aligned}$$

and by lemma 5.2:

$$Z_{k,p+1} = \sum_{x=0}^{p+1}(-1)^x \binom{p+1}{x} 2^{\binom{k-x}{2}}$$

□

Hence the number of representations of a target concept $t \epsilon M_{N,k}$ can be stated:

Proposition 5.4 Representations of $t \epsilon M_{N,k}$. *The number N_t of case-based representations $\langle CB, \sigma_{\overline{w}_t}\rangle$ s.t. $CB \subseteq t$ and $h_{\langle CB, \sigma_{\overline{w}_t}\rangle} = t$ for some function $t \epsilon M_{N,k}$ is defined as follows:*

$$N_t = (2^{2^{N-k}} - 1).2^{\sum_{j>2}^{k} \binom{k}{j}} . \sum_{i=0}^{k} \binom{k}{i} Z_{k,i}$$

where:

$$Z_{k,i} = \sum_{x=0}^{i}(-1)^x . \binom{i}{x} .2^{\binom{k-x}{2}}$$

Proof: First consider how many subsets D' of the partition $\langle D_N, \approx_{\sigma_{\overline{w}_t}} \rangle$ there are s.t. if the intersections of those D' with the descriptions of the exemplars in a case-base $CB \subseteq t$ are non-empty then the function represented by $\langle CB, \sigma_{\overline{w}_t} \rangle$ is precisely t. As above for some case-base CB s.t. $h_{\langle CB, \sigma_{\overline{w}_t} \rangle} = t$, let i be the number of elements in D_1 which are empty. For a particular value of i, the D' containing the positive instances of t must be non-empty, there are $\left(\begin{array}{c} k \\ i \end{array} \right)$ ways of choosing $k - i$ elements of D_1 to be non-empty, there are $Z_{k,i}$ ways of choosing sufficient elements from D_2, and the remaining $\sum_{j>2}^{k} \left(\begin{array}{c} k \\ j \end{array} \right)$ elements of D may be chosen freely. Hence there are $2^{\sum_{j>2}^{k} \left(\begin{array}{c} k \\ j \end{array} \right)} \cdot \sum_{i=0}^{k} \left(\begin{array}{c} k \\ i \end{array} \right) Z_{k,i}$ ways of choosing elements of D to be non-empty. Finally, each of these non-empty equivalence classes contains 2^{N-k} descriptions, giving $(2^{2^{N-k}} - 1)$ ways of choosing exemplars from each equivalence class to make that class non-empty. □

‘Canonical’ Case-Bases and Membership of H_t

The ‘counting’ result of Proposition 5.4 follows immediately from the characterisation of a case-base representing the target concept t given in Proposition 5.1. In addition, the form of the condition in Proposition 5.1 shows that $\langle CB, \sigma_{\overline{w}_t} \rangle$ is a *stable* representation for a target concept $t \in M_{N,k}$, in that all case-bases CB' derived from CB by the addition of exemplars for t ($CB \subseteq CB' \subseteq t$) are also representations for t with respect to $\sigma_{\overline{w}_t}$. This is clearly the case, since CB must contain the exemplars required by proposition 5.1 (necessity), and any $CB' \supseteq CB$ will also contain those exemplars, implying that CB' also represents t (sufficiency). However, this will not be true in the case of case-based representations using the *unweighted* feature count σ_H, nor will it be true in general of other formulations of case-based learning, e.g. [Glo95] [SDHK95].

The ‘stability’ of representations of target concepts with respect to $\sigma_{\overline{w}_t}$ is one corollary of the simplified semantics of the case-based representation in this special case. Another corollary is that for $CB2$-representable functions in H_t (equation (65)) other than t, there is a unique ‘maximal’ case-base that can be computed from the boolean representation of the functions. This gives a straightforward way of determining membership of H_t. Specifically, the results below will demonstrate:

- Firstly, a *canonical* case-base $CB_{(f,t)}$ can be defined such that if a function $f \in B_N$ is $CB2$-representable with respect to a function $t \in M_N$ ($f \in H_t$), then $\langle CB_{(f,t)}, \sigma_{\overline{w}_t} \rangle$ will be a representation for f (Definition 5.4 and Proposition 5.5). This gives a straightforward way, given a boolean representation of a function $f \in B_N$, of determining whether the function is $CB2$-representable w.r.t. to some $t \in M_N$. This is in contrast to the problem of determining whether some function $f \in B_N$ has a case-based representation with respect to the unweighted similarity measure σ_H and some target concept, for which we have no better method than systematically considering all possible case-bases.

- Secondly, these canonical case-bases are maximal, in that all case-bases $CB \subseteq t$ s.t. $h_{\langle CB, \sigma_{\overline{w}_t} \rangle} = f$ are subsets of the canonical case-base $CB_{(f,t)}$ (Also Proposition 5.5).

- Finally, these canonical case-bases are well-behaved in the sense that, given a function $f \in H_t$ which is represented by $CB_{(f,t)}$, then for any function f' which generalises f ($f(d) = 1 \rightarrow f'(d) = 1$) and which is also in H_t, the canonical case-base $CB_{(f',t)}$ will be a subset of $CB_{(f,t)}$ (Proposition 5.8).

Definition 5.4 Canonical Case-Base. *The canonical (‘maximal’) case-base $CB_{(f,t)}$ representing a function $f \in H_t$ with respect to some target function $t \in M_N$ is the case-base containing a positive exemplar for each positive instance of the target function t, and a negative exemplar for each negative instance of*

t, so long as that negative exemplar causes no positive instance of f to be misclassified.

$$\forall d \,\epsilon\, D_N \cdot (t(d) = 1 \wedge \exists d' \,\epsilon\, D_N \cdot f(d') = 1) \leftrightarrow (d,1) \,\epsilon\, CB_{(f,t)} \tag{75}$$

$$\forall d \,\epsilon\, D_N \cdot \\ (t(d) = 0 \wedge [\forall d' \,\epsilon\, D_N \cdot f(d') = 1 \rightarrow \exists d'' \,\epsilon\, D_N \cdot t(d'') = 1 \wedge \sigma_{\overline{w}_t}(d', d) < \sigma_{\overline{w}_t}(d', d'')]) \leftrightarrow (d,0) \,\epsilon\, CB_{(f,t)} \tag{76}$$

Proposition 5.5 *For any target concept $t \,\epsilon\, M_N$, then for a case-base $CB \subseteq t$ let f denote the function represented by CB, $h_{\langle CB, \sigma_{\overline{w}_t}\rangle}$. Then:*

1. *f will also be represented by the canonical representation $\langle CB_{(f,t)}, \sigma_{\overline{w}_t}\rangle$.*
2. *CB will also be a subset of the canonical case-base $CB_{(f,t)}$.*

Proof: Assume that there is some case-base CB extensible to t that represents $f \,\epsilon\, B_N$ with respect to the measure $\sigma_{\overline{w}_t}$.

$$CB \subseteq t \wedge h_{\langle CB, \sigma_{\overline{w}_t}\rangle} = f \tag{77}$$

Define a case-base CB_{max} as follows:

$$CB \subseteq CB_{max} \subseteq t \tag{78}$$

$$h_{\langle CB_{max}, \sigma_{\overline{w}_t}\rangle} = f \tag{79}$$

and

$$|CB_{max}| \geq |CB''| \tag{80}$$

where CB'' is any case-base satisfying:

$$CB \subseteq CB'' \subseteq t \tag{81}$$

$$h_{\langle CB'', \sigma_{\overline{w}_t}\rangle} = f \tag{82}$$

The proof below must cover the following issues:

1. It must be shown that there is exactly one case-base satisfying the equations for CB_{max}.
2. It must then be shown that this maximal extension of the representing case-base is the canonical case-base $CB_{(f,t)}$ of definition 5.4

1) CB_{max} is uniquely defined

There will always be at least one such case-base (if CB cannot be extended without changing the representation, then $CB_{max} = CB$). It must also be shown that there is no more than one maximal extension. For a proof by contradiction, assume the contrary, that there are two distinct case-bases CB^1_{max} and CB^2_{max} satisfying equations (78) to (80). By equation (80) these two case-bases must have the same cardinality. Since they are in addition distinct, there must be at least one exemplar in CB^1_{max} which is in neither CB nor CB^2_{max}. Now this exemplar must be either positive or negative. In the case of a positive exemplar $(d^+_{new}, 1)$, consider the case-base $CB^2_{max} \cup \{(d^+_{new}, 1)\}$. CB^2_{max} must already contain at least one positive exemplar. If not, f would have no positive instances, and CB^1_{max} could have no positive exemplar $(d^+_{new}, 1)$. Since the use of the weighted measure $\sigma_{\overline{w}_t}$ makes all positive instances of t equivalent for the purposes of classification, then $CB^2_{max} \cup \{(d^+_{new}, 1)\}$ must represent exactly the same function as CB^2_{max}, namely f. Hence, neither CB^1_{max} nor CB^2_{max} is maximal, contrary to our assumption.

On the other hand, assume that CB^1_{max} contains an additional negative exemplar $(d^-_{new}, 0)$, and consider the case-base $CB^2_{max} \cup \{(d^-_{new}, 0)\}$. It is clear that adding a negative exemplar to a case-base will not add any new positive instances to the function represented, i.e. the positive instances of $h_{\langle CB^2_{max}\cup\{(d^-_{new},0)\},\sigma_{\overline{w}_t}\rangle}$ will be a subset of the positive instances of $h_{\langle CB^2_{max},\sigma_{\overline{w}_t}\rangle}$. It must only be shown that there are no positive instances of $h_{\langle CB^2_{max},\sigma_{\overline{w}_t}\rangle}$ which will become negative by the addition of $(d^-_{new}, 0)$ to CB^2_{max}. Since CB^1_{max} (which contains $(d^-_{new}, 0)$) represents f, then for any $d \,\epsilon\, D_N$ which is a positive instance of f:

$$\forall d' \,\epsilon\, D_N \cdot t(d') = 1 \rightarrow \sigma_{\overline{w}_t}(d, d') > \sigma_{\overline{w}_t}(d, d^-_{new}) \tag{83}$$

Therefore, since $\langle CB^2_{max}, \sigma_{\overline{w}_t}\rangle$ is also a representation for f, for any positive instance d of $h_{\langle CB^2_{max},\sigma_{\overline{w}_t}\rangle}$, the positive instances of t are strictly more similar to d than d is to d^-_{new}. Therefore its classification will not be changed by the addition of $(d^-_{new}, 0)$ to CB^2_{max}. So, both $(d^-_{new}, 0) \not\epsilon\, CB^2_{max}$ and also $h_{\langle CB^2_{max}\cup\{(d^-_{new},0)\},\sigma_{\overline{w}_t}\rangle} = f$. Hence again neither CB^1_{max} nor CB^2_{max} is maximal, contradicting the original assumption. So by contradiction there can be no more than one maximal case-base.

2) $CB_{max} = CB_{(f,t)}$

If it can be shown that $CB_{max} = CB_{(f,t)}$, then by equation (79) we have the result that $\langle CB_{(f,t)}, \sigma_{\overline{w}_t}\rangle$ is a representation for f and by equation (78) the result that the canonical case-base $CB_{(f,t)}$ includes all representing case-bases CB.

Assume that there is a positive exemplar $(d_{pos}, 1) \,\epsilon\, CB_{max}$. Now $CB_{max} \subseteq t$, so immediately $t(d_{pos}) = 1$. Since in addition there will be at least one point in the example space classified positively (the description d_{pos}), we have:

$$\forall d \,\epsilon\, D_N \cdot (d, 1) \,\epsilon\, CB_{max} \rightarrow (t(d) = 1 \wedge \exists d' \,\epsilon\, D_N \cdot f(d') = 1) \tag{84}$$

Assume instead that a description d is a positive instance of t and that f has at least one positive instance. Finally assume also that $(d, 1)$ is not a positive exemplar in CB_{max}. Consider the case-base $CB' = CB_{max} \cup \{(d, 1)\}$. Clearly $CB_{max} \subseteq CB' \subseteq t$. Note in addition that the similarity of any description $d' \,\epsilon\, D_N$ to the 'missing' description d will be the same as the similarity of d' to any other positive instance of t. Thus so long as CB_{max} contains *some* positive exemplars (which must be the case given that f has at least one positive instance), the function represented by CB' will be exactly that represented by CB_{max}. Hence $h_{\langle CB',\sigma_{\overline{w}_t}\rangle} = h_{\langle CB_{max},\sigma_{\overline{w}_t}\rangle} = f$, and equation (80) is contradicted. Thus:

$$\forall d \,\epsilon\, D_N \cdot (t(d) = 1 \wedge \exists d' \,\epsilon\, D_N \cdot f(d') = 1) \rightarrow (d, 1) \,\epsilon\, CB_{max} \tag{85}$$

Assume now that there exists some negative exemplar $(d_{neg}, 0)$ in CB_{max}. Again, by equation (78), we infer $t(d_{neg}) = 0$ and

$$\forall d \,\epsilon\, D_N \cdot (d, 0) \,\epsilon\, CB_{max} \rightarrow t(d) = 0 \tag{86}$$

Still assuming that $(d_{neg}, 0)$ is an exemplar in CB_{max}, from the fact that $\langle CB_{max}, \sigma_{\overline{w}_t}\rangle$ is a representation of f, then it is clear that any positive instance of f must be more similar to the positive exemplars in CB_{max} than to d_{neg}. Since any positive exemplar in CB_{max} is also a positive instance of t, then also:

$$\forall d \,\epsilon\, D_N \cdot (d, 0) \,\epsilon\, CB_{max} \rightarrow [\forall d' \,\epsilon\, D_N \cdot f(d') = 1 \rightarrow \exists d'' \,\epsilon\, D_N \cdot t(d'') = 1 \wedge \sigma_{\overline{w}_t}(d', d) < \sigma_{\overline{w}_t}(d', d'')] \tag{87}$$

Finally assume that for some negative instance of the target concept $d_{neg} \,\epsilon\, D_N$, then the statement holds as above that any positive instance of f must be more similar to the positive instances of t than to d_{neg}, i.e. it is assumed

$$t(d_{neg}) = 0 \tag{88}$$

$$\forall d' \,\epsilon\, D_N \cdot f(d') = 1 \rightarrow \exists d'' \,\epsilon\, D_N \cdot t(d'') = 1 \wedge \sigma_{\overline{w}_t}(d', d_{neg}) < \sigma_{\overline{w}_t}(d', d'') \tag{89}$$

Assume that $(d_{neg}, 0)$ is not an exemplar in CB_{max}, and consider the case-base $CB' = CB_{max} \cup \{(d_{neg}, 0)\}$. Since $t(d_{neg}) = 0$ then $CB' \subseteq t$ still. As noted above, adding a negative exemplar to the case-base can only result in points in the example space being re-classified from positive to negative, so $h_{\langle CB',\sigma_{\overline{w}_t}\rangle}$ will differ from $f = h_{\langle CB_{max},\sigma_{\overline{w}_t}\rangle}$ only on false negatives. Equation (89), however, guarantees that for any positive instance of f, there is a positive instance of t which is more similar to that instance than is d_{neg}. Since equation (85) indicates that the positive instance of t required by (89) will indeed be an exemplar of the case-base (as long as there is at least one description satisfying f), then it is clear that no positive instance of f will be misclassified by $h_{\langle CB',\sigma_{\overline{w}_t}\rangle}$. Hence equation (80) is contradicted and we infer that $(d_{neg}, 0)$ must be in CB_{max}. Hence:

$$\forall d \,\epsilon\, D_N \cdot$$
$$(t(d) = 0 \wedge [\forall d' \,\epsilon\, D_N \cdot f(d') = 1 \rightarrow \exists d'' \,\epsilon\, D_N \cdot t(d'') = 1 \wedge \sigma_{\overline{w}_t}(d', d) < \sigma_{\overline{w}_t}(d', d'')]) \rightarrow (d, 0) \,\epsilon\, CB_{max} \tag{90}$$

Thus equations (84), (85), (86), (87) and (90) are precisely the definition of $CB_{(f,t)}$, and $CB_{max} = CB_{(f,t)}$. □

Example 5.6 Is the function $u_1u_2 + u_1u_3u_4 + u_2u_3u_4$ $CB2$-representable with respect to the target concept $u_1u_2u_3u_4$ on the space D_4 ?

Let f be the function represented by $u_1u_2 + u_1u_3u_4 + u_2u_3u_4$ and t be the function represented by $u_1u_2u_3u_4$. From definition 5.4, $CB_{(f,t)}$ contains positive exemplars for precisely those descriptions which are positive instances of t. Hence there is a single positive exemplar $(1111, 1)$. Also, there is a negative exemplar for every description which is less similar to any positive instance of f than the positive instance of t. Consider those descriptions which satisfy the disjunct u_1u_2, namely 1111, 1110, 1101 and 1100. For each such disjunct, it is only necessary to consider the minimal true vector (1100), since any description which is less similar to 1100 than 1100 is to the positive exemplar 1111 will also be sufficiently dissimilar to any of the other descriptions satisfying u_1u_2. Therefore the similarity of any negative exemplar in $CB_{(f,t)}$ to the description 1100 must be $< \frac{1}{2}$, which is equivalent to requiring a similarity of at least $\frac{3}{4}$ to 0011 which is the complement of the minimal true vector 1100. Hence, considering the other two disjuncts in the same way, the negative exemplars in $CB_{(f,t)}$ are precisely those descriptions d_{neg} satisfying:

$$\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_H(d_{neg}, 0011) \geq \frac{3}{4} \tag{91}$$
$$\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_H(d_{neg}, 0100) \geq \frac{2}{4} \tag{92}$$
$$\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_H(d_{neg}, 1000) \geq \frac{2}{4} \tag{93}$$

(Note that $\sigma_{\overline{w}_t} = \sigma_H$ for a ‘fully defined’ target concept such as $t = u_1u_2u_3u_4$.) Thus $CB_{(f,t)}$ contains the positive exemplar $(1111, 1)$ and the negative exemplars $(0001, 0)$ and $(0010, 0)$.

Proposition 5.5 indicates that if f is $CB2$-representable with respect to t at all, then $h_{\langle CB_{(f,t)},\sigma_H\rangle} = f$.

The boolean representation of the function represented by $\langle CB_{(f,t)}, \sigma_H\rangle$ is:

$$\Delta_{\sigma_H}(1111, 0001).\Delta_{\sigma_H}(1111, 0010)$$

where, as in Example 3.4:

$$\Delta_{\sigma_H}(d_{pos}, d_{neg}) = \bigvee\{U \subseteq (U_{d_{pos}} \setminus U_{d_{neg}}) | \#U = \left\lfloor \frac{\#(U_{d_{pos}} \setminus U_{d_{neg}})}{2} \right\rfloor + 1\}$$

Hence:

$$\begin{aligned} & h_{\langle CB_{(f,t)},\sigma_H\rangle} \\ = \;& (u_1u_2 + u_1u_3 + u_2u_3)(u_1u_2 + u_1u_4 + u_2u_4) \\ = \;& u_1u_2 + u_1u_2u_4 + u_1u_2u_4 + u_1u_2u_3 + u_1u_3u_4 + u_1u_2u_3u_4 + u_1u_2u_4 + u_1u_2u_3u_4 + u_2u_3u_4 \\ = \;& u_1u_2 + u_1u_3u_4 + u_2u_3u_4 \end{aligned}$$

Hence the function $f = u_1u_2 + u_1u_3u_4 + u_2u_3u_4$ *is* $CB2$-representable with respect to $t = u_1u_2u_3u_4$. □

Example 5.7 Is the function $u_1u_2 + u_1u_3 + u_1u_4$ $CB2$-representable with respect to the target concept $u_1u_2u_3u_4$ on the space D_4 ?

Repeating the previous argument, let f be the function represented by $u_1u_2 + u_1u_3 + u_1u_4$ and let t be the target concept $u_1u_2u_3u_4$. Then the canonical case-base $CB_{(f,t)}$ contains the positive exemplar $(1111, 1)$ and those negative exemplars satisfying:

$$\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_H(d_{neg}, 0011) \geq \frac{3}{4} \tag{94}$$

$$\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_H(d_{neg}, 0101) \geq \frac{3}{4} \tag{95}$$

$$\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_H(d_{neg}, 0110) \geq \frac{3}{4} \tag{96}$$

So $CB_{(f,t)} = \{(1111, 1), (0111, 0)\}$ and $h_{\langle CB_{(f,t)}, \sigma_H \rangle} = \Delta_{\sigma_H}(1111, 0111) = u_1$. Hence $u_1u_2 + u_1u_3 + u_1u_4$ is not $CB2$-representable with respect to the target concept $u_1u_2u_3u_4$. □

Finally, as noted above, these canonical representations have the property that given two functions $f, f' \,\epsilon\, H_t$ s.t. f' generalises f, then the canonical case-base for f' will be a *subset* of the canonical case-base for f.

Proposition 5.8 *Given a function $f \,\epsilon\, B_N$ that is $CB2$-representable with respect to a target concept $t \,\epsilon\, M_N$, then any function $f' \,\epsilon\, B_N$ which is a generalisation of f i.e.*

$$\forall d \,\epsilon\, D_N \cdot f(d) = 1 \rightarrow f'(d) = 1$$

and is $CB2$-representable w.r.t. t, will be represented by a canonical case-base $CB_{(f',t)}$ representing f will be a subset of $CB_{(f,t)}$.

$$CB_{(f',t)} \subseteq CB_{(f,t)}$$

Proof: Given $f, f' \,\epsilon\, B_N$ where f' generalises f ($\forall d \,\epsilon\, D_N \cdot f(d) = 1 \rightarrow f'(d) = 1$), then $CB_{(f,t)} \triangle CB_{(f,t)}$ will only contain some negative exemplars for t which are present in $CB_{(f,t)}$ but not in $CB_{(f',t)}$. This is clear from the definition of the canonical case-base (Definition 5.4); adding extra positive instances to the represented function clearly requires the removal of negative instances from the canonical case-base. □

Characterisation of $CB2$-representable functions

Finally, it has been possible to give some necessary conditions for functions that are $CB2$-representable with respect to some target concept, although a necessary and sufficient characterisation has proved elusive. The results to be presented are:

- All functions that are $CB2$-representable can also be represented by boolean expressions in DNF where all disjuncts are subsumed by the monomial expression for the target concept.
- Also, such a boolean expression for a $CB2$-representable function will contain no pairs of disjuncts that are disjoint, having no literals in common.

The first claim is given in Proposition 5.12, the second in Proposition 5.15 below. Proposition 5.12 depends on the results of Propositions 5.10 and 5.11, while Proposition 5.15 builds on 5.13 and 5.14. The expression of these results also depends on some notions of minimal and irreducible boolean representations of functions which are stated immediately below. These are reproduced from [Mur71], but equivalent statements should be found in any work on switching theory.

Definition 5.5 *[Mur71, Defn 2.1.8] If there exists a disjunctive form for a function $f \epsilon B_N$ such that the literal $\overline{u}_i$ does not appear in any term of this form, then f is said to be* **positive** *in u_i, u_i is a* **positive variable** *of f and the disjunctive form is said to be* **positive** *in u_i. Conversely, if a disjunctive form for a function $f \epsilon B_N$ contains no literal u_i, then f is said to be* **negative** *in u_i, u_i is a* **negative variable** *of f and the disjunctive form is said to be* **negative** *in u_i.*

(Note that f is both positive and negative in a variable u_i if neither u_i nor $\overline{u}_i$ appear in a particular disjunctive form for f, or the value of f is otherwise independent of u_i.)

Definition 5.6 *[Mur71, Defn 2.1.8] A function that is positive in all variables is a* **positive function***. A function that is negative in all variables is a* **negative function***.*

Definition 5.7 *[Mur71, Defn 2.1.8] If f or a disjunctive form representing f is either positive or negative in a variable u_i, then f is said to be* **unate** *in u_i. A function (or a disjunctive form) which is unate in all variables is called a* **unate function** *(or a unate disjunctive form).*

Theorem 5.9 *[Mur71, Thm 2.1.6] A unate function f has exactly one irreducible disjunctive form. Furthermore, the form is unate and consists of all prime implicants of the function.*

In what follows, the functions that are $CB2$-representable with respect to some target concept $t \epsilon M_N$ are characterised in terms of properties of the minimal DNF representation of the function required by Theorem 5.9. When discussing the boolean representation of a function $f \epsilon B_N$, it is this irreducible expression that is intended.

The first result necessary for the proof of Proposition 5.12 simply restates the result of Proposition 3.5 in the case of the weighted similarity measure $\sigma_{\overline{w}_t}$. This is given without proof because of its similarity to the earlier result.

Proposition 5.10 *Consider a case-base CB extensible to some $t \epsilon M_{N,k}$ and a point $d \epsilon D_N$ which is a positive instance of $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}$. Any positive exemplar $(d_1^+, 1) \epsilon CB$ will be maximally similar according to $\sigma_{\overline{w}_t}$ w.r.t. all other exemplars in CB:*

$$\forall (d', n) \epsilon CB \cdot \sigma_{\overline{w}_t}(d, d_1^+) \geq \sigma_{\overline{w}_t}(d, d')$$

Then, any instance which lies on a shortest path (reckoned only with respect to bits of the representation relevant to t) through the example space between d and d_1^+, i.e. any point $d^* \epsilon D_N$ s.t.

$$\forall 1 \leq i \leq N \cdot ((w_t)_i > 0 \wedge (d^*)_i \neq (d_1^+)_i) \to (d)_i \neq (d_1^+)_i$$

is also a positive instance of $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}$:

$$h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(d^*) = 1$$

Since all positive exemplars in the case-base are equivalent for the purposes of classification due to the weighting of $\sigma_{\overline{w}_t}$, this result indicates that all functions that are $CB2$-representable can be expressed as a sum of terms where each term is subsumed by the monomial target expression; this is the result expressed in Proposition 5.12. Proposition 5.11 serves as a lemma in the proof of Proposition 5.12.

Definition 5.8 *A function $f_1 \epsilon B_N$* **agrees** *with a function $f_2 \epsilon B_N$ iff f_1 is positive on any variable u_i on which f_2 is positive and f_1 is negative on any variable u_i on which f_2 is negative.*

Proposition 5.11 *Given a function $t \epsilon M_N$ and a function $f \epsilon B_N$, then f will agree with t, provided the following condition holds. Let $d \epsilon D_N$ be a positive instance of f. Let d^+ be the 'maximal true vector' for t, i.e. the vector which has value 1 at every bit except where the monomial expression for t contains a negative literal $\overline{u}_i$. Then any vector d' on a direct path from d to d^+, i.e. any $d' \epsilon D_N$ s.t. d' agrees* *with d^+ wherever d agrees with d', is also a positive instance of f. That is, a function f agrees with t provided that:*

$$\forall d \,\epsilon\, D_N \cdot f(d) = 1 \rightarrow \forall d' \,\epsilon\, D_N \cdot [\forall 1 \leq i \leq N \cdot (d)_i = (d^+)_i \rightarrow (d')_i = (d^+)_i] \rightarrow f(d') = 1$$

where d^+ is the maximal true vector satisfying t.

Proof: For any positive instance $d \,\epsilon\, D_N$ s.t. $f(d) = 1$ assume that there is a monomial expression U s.t.:

1. d is a positive instance of h_U
2. all other positive instances of h_U are positive instances of f
3. h_U agrees with t

Consider the DNF formula which is the disjunction of such U over all positive instances of f. For a positive instance of f, there will be at least one disjunct which is true on that instance, so that the whole expression will be true on all that instance. Equally for a negative instance of f, none of the disjuncts U will be true, making the whole expression false. This disjunction is clearly an expression for f, and additionally the expression will contain no negative literal $\overline{u}_i$ if no such literal appears in the monomial expression for t, and will contain no positive literal if no such literal appears in the monomial expression for t, since this will be true of each individual disjunct; f in this case agrees with t.

Thus to show that a function $f \,\epsilon\, B_N$ agrees with $t \,\epsilon\, M_N$, it will be sufficient to show that for any positive instance of f, there is a monomial U satisfying the three conditions above. Assume therefore that the stated precondition holds for some $d \,\epsilon\, D_N$ where $f(d) = 1$.

$$\forall d \,\epsilon\, D_N \cdot f(d) = 1 \rightarrow \forall d' \,\epsilon\, D_N \cdot [\forall 1 \leq i \leq N \cdot (d)_i = (d^+)_i \rightarrow (d')_i = (d^+)_i] \rightarrow f(d') = 1 \qquad (97)$$

where d^+ is the maximal true vector satisfying t as above.

Let T be the monomial expression representing the target concept t. Consider the monomial $U \subseteq T$ s.t. a literal from T appears in U iff d and d^+ agree on the corresponding bit of the representation:

$$\forall 1 \leq i \leq N \cdot u_i \,\epsilon\, U \leftrightarrow (u_i \,\epsilon\, T \wedge (d)_i = 1) \qquad (98)$$

$$\forall 1 \leq i \leq N \cdot \overline{u}_i \,\epsilon\, U \leftrightarrow (\overline{u}_i \,\epsilon\, T \wedge (d)_i = 0) \qquad (99)$$

Clearly $h_U(d) = 1$, since $u_i \,\epsilon\, U \rightarrow (d)_i = 1$ and $\overline{u}_i \,\epsilon\, U \rightarrow (d)_i = 0$; the first condition is satisfied immediately. Equally, $u_i \,\epsilon\, U \rightarrow u_i \,\epsilon\, T$ and $\overline{u}_i \,\epsilon\, U \rightarrow \overline{u}_i \,\epsilon\, T$. Thus h_U clearly agrees with $t = h_T$ according to definition 5.8, satisfying the third condition also.

It only remains to show that all other positive instances of h_U, $d' \,\epsilon\, D_N$ s.t. $h_U(d') = 1$ are also positive instances of f. Assume therefore that some description $d' \,\epsilon\, D_N$ is a positive instance of h_U. Consider the description d'' s.t. d'' and d' agree on those bits which irrelevant to the target concept $t = h_T$ $(\{u_1, \overline{u}_i\} \cap T = \emptyset)$ and d'' & d agree on all bits which *are* relevant to t. So for irrelevant bits we have $(d'')_i = (d')_i$ and:

$$\forall 1 \leq i \leq N \cdot \{u_1, \overline{u}_i\} \cap T = \emptyset \rightarrow (d'')_i = (d^+)_i \rightarrow (d')_i = (d^+)_i \qquad (100)$$

Consider also that d' is a positive instance of h_U, as is d. So d' and d agree on any bit of the representation constrained by U; $\{u_i, \overline{u}_i\} \cap U \neq \emptyset \rightarrow (d')_i = (d)_i$. Since U corresponds to precisely those bits of the representation which are relevant to the definition of the target concept and on which d and d^+ agree, for any bit of the representation relevant to t, $(d)_i = (d^+)_i \rightarrow (d')_i = (d^+)_i$. But since d'' and d agree on all bits constrained by the target monomial U, then also:

$$\forall 1 \leq i \leq N \cdot \{u_1, \overline{u}_i\} \cap T \neq \emptyset \rightarrow (d'')_i = (d^+)_i \rightarrow (d')_i = (d^+)_i \qquad (101)$$

So for any bit of the representation, d' must agree with d^+ whenever d'' and d^+ agree. Since the assumed precondition (97) also applies to d'', then $f(d') = 1$, so that $h_U(d') = 1 \rightarrow f(d') = 1$ as required in the second condition. □

Proposition 5.12 *All functions $f \epsilon H_t$, i.e. those functions $f \epsilon B_N$ which are $CB2$-representable with respect to a target concept $t \epsilon M_N$, are unate functions, and further, the represented function f will always agree with the target concept t according to definition 5.8.*

Proof: Take a function f with a case-based representation $\langle CB, \sigma_{\overline{w}_t} \rangle$ derived from a target $t \epsilon M_N$. For a proof by contradiction, we must assume either that f is not unate, or that t is positive in some variable in which f is not, or that t is negative in some variable and f is not.

Suppose f is not unate. Then there is some variable u_i s.t. for all forms of the function f both u_i and $\overline{u}_i$ appear in the expression. Thus whether t is positive or negative in that variable, f will be neither positive nor negative. Hence there are two cases to consider, one where t is positive in some variable but f is not, and one where t is negative in some variable but f is not; if f is not unate in some variable then one or other of these conditions will be implied.

In the case where t is positive in some variable but f is not (i.e. the monomial expression for t does not contain the literal $\overline{u}_i$, but $\overline{u}_i$ appears in all boolean expressions for f) then Proposition 5.11 indicates that there must be a positive instance d of f s.t. there is a description d' lying on a direct path between d and the description d^+ which is the maximum true vector for t, and d' is a negative instance of f.

$$\exists d \epsilon D_N \cdot f(d) = 1 \wedge \exists d' \epsilon D_N \cdot [\forall 1 \leq i \leq N \cdot (d)_i = (d^+)_i \rightarrow (d')_i = (d^+)_i] \wedge f(d') = 0 \qquad (102)$$

Now since this d is a positive instance of f, in the case-based representation $\langle CB, \sigma_{\overline{w}_t} \rangle$ there must be a positive exemplar $(d_{pos}, 1) \epsilon CB$ where d_{pos} is maximally similar to d with respect to the rest of the case-base. Since this is a positive exemplar, then $t(d_{pos}) = 1$ and hence d_{pos} must agree with the maximal true vector d^+ on all bits of the representation relevant to t. So any relevant bit of the representation that d and d_{pos} agree on, d' and d_{pos} will also agree on (equation (102)). Thus d' and d_{pos} will disagree on a relevant bit of the representation only if d and d_{pos} also disagree. Hence by proposition 5.10 we have

$$h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(d') = 1 \qquad (103)$$

for any $d' \epsilon D_N$ satisfying $\forall 1 \leq i \leq N \cdot (d)_i = (d^+)_i \rightarrow (d')_i = (d^+)_i$, contradicting equation (102).

The argument may be repeated in the case where t is negative in some variable and f is not. Hence by contradiction in each case, it is concluded that f must be unate and that f agrees with the target concept t according to definition 5.8. □

Proposition 5.12 shows that all $CB2$-representable functions are unate and must agree with their target concept. That is, a positive (or negated) literal can appear in the irreducible disjunctive form of f only if the same positive (or negated) literal appears in the monomial expression of the target concept t. The corollary of this is that each term of the irreducible disjunctive form of f must be subsumed by the expression for the target concept, as stated above.

The second necessary property of $CB2$-representable functions is that no disjunctive form of the function may contain disjoint disjuncts. The argument proceeds by showing that no $CB2$-representable function can represent a function which has two positive instances which are maximally separate in the example space $\sigma_{\overline{w}_t}(d_1, d_2) = 0$ (Proposition 5.13). On the other hand, a function represented by an expression containing disjoint disjuncts implies precisely this condition (Proposition 5.14) so that it is concluded that no such function is $CB2$-representable (Proposition 5.15).

Proposition 5.13 *For all values $1 \leq k \leq N$, for any target concept $t \epsilon M_{N,k}$ then there is no function that is $CB2$-representable with respect to t (except the all-1 function f_1^N) that will classify positively two instances $d, d' \epsilon D_N$ which are 'opposite' in the example space (i.e. are as dissimilar as possible and have zero similarity).*

$$\begin{aligned} &\forall 1 \leq k \leq N \cdot \forall t \epsilon M_{N,k} \cdot \forall f \epsilon H_t \cdot f \neq f_1^N \rightarrow \\ &\qquad \forall d_1, d_2 \epsilon D_N \cdot [\sigma_{\overline{w}_t}(d_1, d_2) = 0 \rightarrow (f(d_1) = 0 \vee f(d_2) = 0)] \end{aligned}$$

Proof: Take some case-base CB extensible to a target concept $t \epsilon M_{N,k}$ which contains at least one negative exemplar, and consider two descriptions $d_1, d_2 \epsilon D_N$. Let the similarity of σ_1 denote the similarity of d_1 to the positive instances of the target concept, similarly σ_2.

$$\forall d \epsilon D_N \cdot t(d) = 1 \rightarrow \sigma_{\overline{w}_t}(d, d_1) = \sigma_1$$
$$\forall d \epsilon D_N \cdot t(d) = 1 \rightarrow \sigma_{\overline{w}_t}(d, d_2) = \sigma_2$$

Assume that $\sigma_{\overline{w}_t}(d_1, d_2) = 0$; this will be the case iff d_1 and d_2 disagree on all relevant bits (those with non-zero weighting in $\overline{w}_t$). Hence a description d agrees with d_1 on some relevant bit i iff d disagrees with d_2 on that bit. If δ_1 is the number of relevant bits on which d and d_1 differ, δ_2 is the number on which d and d_2 differ, then we have $\delta_1 + \delta_2 = k$, $1 = \frac{k-\delta_1}{k} + \frac{k-\delta_2}{k}$ and

$$\forall d \epsilon D_N \cdot \sigma_{\overline{w}_t}(d, d_1) + \sigma_{\overline{w}_t}(d, d_2) = 1 \tag{104}$$

Thus immediately $\sigma_1 + \sigma_2 = 1$ since $\sigma_{\overline{w}_t}(d_{pos}, d_1) + \sigma_{\overline{w}_t}(d_{pos}, d_2) = 1$ for any positive instance d_{pos} of the target concept t. Take then some point $d \epsilon D_N$ such that $\sigma_{\overline{w}_t}(d, d_1) < \sigma_1$. Since equation (104) also applies to this instance d we have:

$$\sigma_{\overline{w}_t}(d, d_1) < \sigma_1 \tag{105}$$
$$\sigma_{\overline{w}_t}(d, d_1) + \sigma_{\overline{w}_t}(d, d_2) = 1 \tag{106}$$
$$\sigma_1 + \sigma_2 = 1 \tag{107}$$

giving $1 - \sigma_2 > 1 - \sigma_{\overline{w}_t}(d, d_2)$, and hence $\sigma_{\overline{w}_t}(d, d_2) > \sigma_2$. Thus it has been shown $\sigma_{\overline{w}_t}(d, d_1) < \sigma_1 \rightarrow \sigma_{\overline{w}_t}(d, d_2) > \sigma_2$ which in turn implies $\sigma_{\overline{w}_t}(d, d_2) \geq \sigma_2$. Thus

$$\forall d \epsilon D_N \cdot \sigma_{\overline{w}_t}(d, d_1) \geq \sigma_1 \vee \sigma_{\overline{w}_t}(d, d_2) \geq \sigma_2 \tag{108}$$

Since σ_i is the similarity of the description d_i to any positive instance of the target concept, then this requires that at any possible negative exemplar in the case-base will be at least as close to one of d_1 or d_2 as any possible positive exemplar is. Thus d_1 and d_2 cannot both be positive instances of the represented function unless the case-base contains no negative exemplars, in which case the function represented will be either all-1 or all-0. □

Proposition 5.14 *For any unate function $f \epsilon B_N$ there are two descriptions $d_1, d_2 \epsilon D_N$ s.t. $f(d_1) = f(d_2) = 1$ and in addition $\sigma_H(d_1, d_2) = 0$, iff the unique irreducible disjunctive form for f contains a pair of disjuncts that are disjoint i.e. which have no literal u_i or $\overline{u}_i$ in common.*

Proof: **1) Only if.** For some unate function $f \epsilon B_N$ assume a pair of positive instances of f, d_1 and d_2, such that:

$$\sigma_H(d_1, d_2) = 0$$

Let Φ be the unique, irreducible disjunctive form representing f, and let Π be the union over all the literals appearing in Φ. Let d^+ be the description in D_N which has value 1 on all bits i of the representation s.t. $u_i \epsilon \Pi$, 0 on all bits i s.t. $\overline{u}_i \epsilon \Pi$ and value 1 if $\{u_i, \overline{u}_i\} \cap \Pi = \emptyset$. Note that such a description can only be defined for a unate function f; otherwise Π will in general contain both u_i and $\overline{u}_i$ for some bit i. Let U_1 be the most specific monomial which includes both d^+ and d_1 and U_2 be the most specific monomial including d^+ and d_2. Since d_1 and d_2 agree on no bit of the representation, then U_1 and U_2 must be disjoint.

Now the set of positive instances of the function represented by either U_i, $i \epsilon \{1, 2\}$ will be a subset of the positive instances of f, demonstrated as follows. Assume that there is an example d which is a positive instance of h_{U_i} so that $\forall 1 \leq j \leq N \cdot (d_i)_j = (d^+)_j \rightarrow (d)_j = (d^+)_j$, or any bit on which d_i agrees with d^+, then also d agrees with d^+. By the contrapositive any bit of the representation on which d disagrees with d^+, d_i & d^+ will also disagree:

$$\forall 1 \leq j \leq N \cdot (d)_j \neq (d^+)_j \rightarrow (d_i)_j \neq (d^+)_j \tag{109}$$

Assume also that d is a negative instance of f. Hence for all of the disjuncts U^f in Φ, $h_{U^f}(d) = 0$. Now if $h_{U^f}(d) = 0$, then there must be some literal u_i or $\overline{u}_i$ in U_f s.t. $(u_i \,\epsilon\, U_f \wedge (d)_i = 0) \vee (\overline{u}_i \,\epsilon\, U_f \wedge (d)_i = 1)$. But, by definition, this literal would appear in the union of literals Π which defines d^+ so that $(u_i \,\epsilon\, U_f \rightarrow (d^+)_i = 1) \wedge (\overline{u}_i \,\epsilon\, U_f \rightarrow (d^+)_i = 0)$. d will disagree with d^+ on that bit, and from equation (109) then d_i also disagrees with d^+. Hence any such disjunct U_f must also be zero on d_i. Thus d_i would be a negative instance of f, which is contradictory. Hence all positive instances of h_{U_i} will also be positive instances of f. Hence there must be a disjunct in Φ which includes U_1, and one which includes U_2. Since these can only be more general than U_1 and U_2, then if U_1 and U_2 are disjoint, then the disjuncts in the formula must also be disjoint.

2) If. Assume that for a unate function $f \,\epsilon\, B_N$, then Φ, the unique irreducible disjunctive form of f, contains a pair of disjoint disjuncts. Let U_1 and U_2 be this pair of disjoint disjuncts in Φ, and consider the descriptions $d_1, d_2 \,\epsilon\, D_N$ defined as follows. For any bit i s.t. $u_i \,\epsilon\, U_1$ or $\overline{u}_i \,\epsilon\, U_2$ let $(d_1)_i = 1$ and $(d_2)_i = 0$. For any bit i s.t. $\overline{u}_i \,\epsilon\, U_1$ or $u_i \,\epsilon\, U_2$ let $(d_1)_i = 0$ and $(d_2)_i = 1$. For all other bits let $(d_1)_i$ be 1 and let $(d_2)_i$ be 0. Clearly $h_{U_1}(d_1) = 1$ and $h_{U_2}(d_2) = 1$, so $f(d_1) = f(d_2) = 1$, and in addition, d_1 and d_2 differ on all bits of the representation, so $\sigma_H(d_1, d_2) = 0$. □

Proposition 5.15 $CB2$**-Representable Functions.** *For any $N \geq 1$ and functions $f \,\epsilon\, B_N, t \,\epsilon\, M_N$, f is $CB2$-representable with respect to t only if*

1. *f is a unate function,* **and**
2. *the unique, irreducible disjunctive form representing f contains no pair of disjuncts that are disjoint.*

Proof: The first condition is stated in proposition 5.12. Assume that a function $f \,\epsilon\, B_N$ has a case-based representation $\langle CB, \sigma_{\overline{w}_t} \rangle$ derived from some monomial $t \,\epsilon\, M_N$ as described above. Assume additionally that CB contains both positive and negative exemplars. (If not then the second condition follows trivially). From proposition 5.13 then, there is no pair of examples which are 'opposite' according to $\sigma_{\overline{w}_t}$ ($\sigma_{\overline{w}_t}(d_1, d_2) = 0$), but which are both positive instances of $f = h_{\langle CB, \sigma_{\overline{w}_t} \rangle}$. Since this is true for *no* pair of examples s.t. $\sigma_{\overline{w}_t}(d_1, d_2) = 0$, then it cannot be true for the special case of two descriptions which differ on *all* bits of the representation ($\sigma_H(d_1, d_2) = 0$). Proposition 5.12 asserts that $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}$ must be a unate function so from proposition 5.14 it follows that the irreducible disjunctive form of f contains no pair of disjuncts U_1 and U_2 which are disjoint. □

5.2 Consistency of $CB2$

In the previous section we have been able to map out some of the properties of the case-based representation for the special case of the weighted similarity measure in a little more detail than in the general case of representations using σ_H. Our aim remains the same, however: to use these results in characterising the sample complexity of our learning 'algorithm' $CB2$. The other preliminary, as in §4, is to establish the consistency of $CB2$ in order to apply the standard learning results for sample complexity based on the cardinality and VC dimension of the hypothesis space. Since the weight vector $\overline{w}$ is defined (Definition 5.2) only in terms of monomial target concepts, the results below only consider the case of $C = M_N$; the resemblance between $CB2$ and $CB1(\sigma)$ means that the proof proceeds in a very similar fashion to theorem 4.3 above.

Proposition 5.16 *$CB2$ is a consistent learning algorithm for M_N.*

Proof: For any target concept $t \,\epsilon\, M_N$, the hypothesis output by $CB2$ on $\overline{s}$, a training sample for t, is $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}$, where $CB = \bigcup_{1 \leq i \leq m}(d_i, b_i)$ and $\sigma_{\overline{w}_t}$ is the weighted similarity measure defined in terms of t (Definition 5.2). Dispensing with the trivial case of $t \,\epsilon\, M_{N,0}$ (since $h_{\{\}}$ has value 1 on all $d \,\epsilon\, D_N$ and thus $\forall d \,\epsilon\, D_N \cdot CB2(\overline{s})(d) = t(d) = 1$ on all non-empty training samples $\overline{s}$), we assume below that $\overline{w}_t$ has at least one non-zero bit. Consider the positive and negative examples in $\overline{s}$ separately. First take

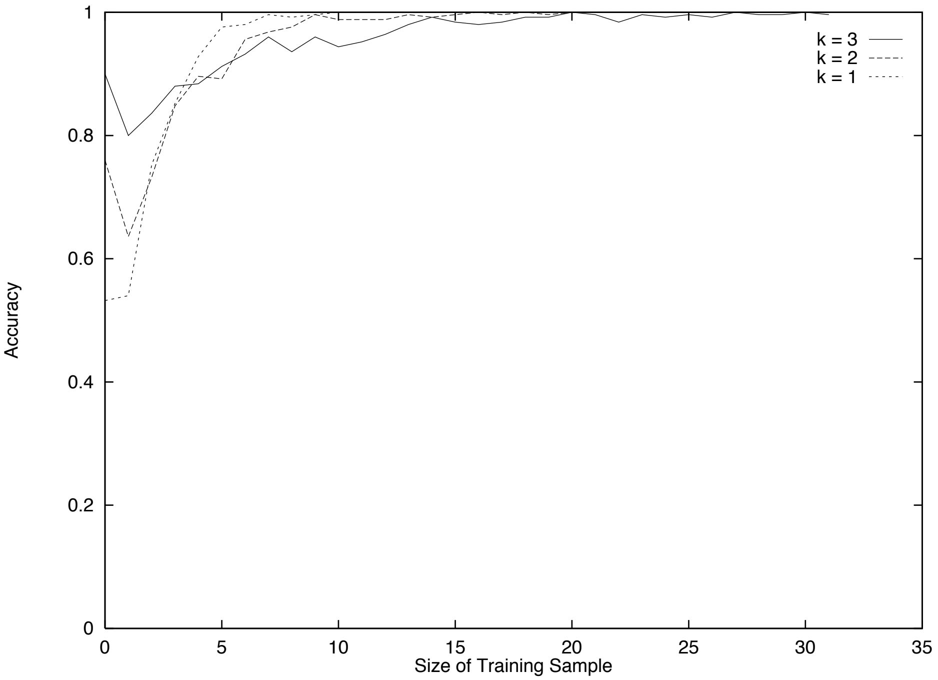

Figure 5: Average Learning Curve For Monomial Functions of k Literals by CB2

some positive example d_i s.t. $b_i = 1$. Since a bit $(w_t)_j$ of the weight vector is > 0 iff the j-th bit of the representation is referred to in the monomial expression defining t, and since any negative exemplar must disagree with d_i on at least one of these defining bits, then clearly:

$$\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_{\overline{w}_t}(d_i, d_{neg}) < \sigma_{\overline{w}_t}(d_i, d_i)$$

Hence $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(d_i) = 1$, by equation (4). Secondly, consider negative examples from the sample, d_i s.t. $b_i = 0$. Again, any positive exemplar $(d_{pos}, 1) \,\epsilon\, CB$, must disagree with d_i on at least one of the defining bits and hence:

$$\forall (d_{pos}, 1) \,\epsilon\, CB \cdot \sigma_{\overline{w}_t}(d_i, d_{pos}) < \sigma_{\overline{w}_t}(d_i, d_i)$$

Hence *a fortiori* $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(d_i) = 0$. Thus for any element of the sample $CB2(\overline{s})(d_i) = b_i$. □

Corollary 5.17 *$CB2$ is a consistent learning algorithm for $M_{N,k}$.*

5.3 Empirical Investigation of $CB2$

The experiments generating figure 4 were repeated in order to measure the learning behaviour of $CB2$ with respect to specific concept spaces. The aim, as before, was to investigate empirically the sample sizes required for reliable hypotheses before addressing formally the question of sample complexity. Figure 5 shows results for 250 training samples for different members of $M_{6,k}$, $1 \leq k \leq 3$, where irrelevant features are weighted 0 and relevant attributes are weighted 1. The figure shows that $CB2$ learns $M_{6,1}$ more quickly than $M_{6,2}$, which is learnt more quickly than $M_{6,3}$. An accuracy of $> 95\%$ is reached on average with a training sample of size 5 for $M_{6,1}$, while the size of training sample needed to reach the same level for $M_{6,2}$ and $M_{6,3}$ is ~ 7 and ~ 10 respectively. This reverses the trend revealed in Figure 4, which showed the learning behaviour of $CB1$ equipped with the unweighted feature count σ_H.

The learning curves in Figure 5 are also a clear improvement over Figure 4; $CB2$ is evidently, on average, much more efficient with respect to sample complexity than $CB1(\sigma_H)$. The extra 'knowledge' or information encoded in the weighted similarity measure in the hypotheses output by $CB2$ has a very marked effect on the learning behaviour. A similar effect is demonstrated over a number of different similarity measures in the learning of concepts defined on ordinal-valued attributes in work by Globig & Wess [GW94].

5.4 Sample Complexity of CB2

Having addressed basic issues of representability and consistency for $CB2$, and explored briefly the average case behaviour by empirical testing, we can now focus on the primary issue of the sample complexity of $CB2$. As before, the sample complexity will be explored by considering the effective hypothesis space of $CB2$ with respect to different classes of target concept. Fortunately, with respect to the case-base part of the hypothesis representation, $CB2$ will have exactly the same behaviour as any $CB1(\sigma)$ in that a 'collect all' strategy is employed in constructing the case-base. Hence proposition 4.1 can be restated in the case of $CB2$:

Proposition 5.18 *A function f is a member of the hypothesis space of $CB2$ with respect to the concept space $C \subseteq B_N$ if and only if there is some target concept $c \in C$ for which there is a case base $CB \subseteq c$ s.t. $h_{\langle CB, \sigma_{\overline{w}_c} \rangle} = f$, where $\sigma_{\overline{w}_c}$ is defined with knowledge of the target concept as in definition 5.2.*

$$\forall C \subseteq B_N \cdot \forall f \in B_N \cdot f \in H_C^{CB2} \leftrightarrow \exists c \in C \cdot \exists CB \subseteq c \cdot h_{\langle CB, \sigma_{\overline{w}_c} \rangle} = f$$

As a result of this, the set of functions that might be output by $CB2$ on a training sample for some concept $t \in M_{N,k}$ is precisely the set of $CB2$-representable functions H_t, and in analogy to equation (17), the effective hypothesis space of $CB2$ with respect to some concept space C can be defined:

$$H_C^{CB2} = \bigcup_{t \in C} H_t \tag{110}$$

where H_t is defined as in Definition 5.3.

Hypothesis Space of $CB2$ increases with k

Figure 4 showed the empirically observed learning curve for $CB1(\sigma_H)$ with respect to target concepts in the space $M_{N,k}$. The graph showed how for such target concepts, the average case 'sample complexity' of $CB1(\sigma_H)$ decreases as k increases. Figure 5 on the other hand shows the opposite trend in the learning behaviour of $CB2$ with respect to targets in $M_{N,k}$; sample complexity increases with k. We consider in this section how this might be reflected in changes to the structure of the hypothesis space of $CB2$ as compared to $CB1(\sigma_H)$. Note from the discussions on representation in section 5.1 that the functions in $H_{M_{N,k}}^{CB2}$ can depend only on variables relevant to the target concept. Thus, in place of Proposition 4.17, it is straightforward to show that $\forall 1 \leq k \leq N \cdot H_{M_{N,k}}^{CB2} \cap M_N = \bigcup_{0 \leq k' \leq k} M_{N,k'}$; only those monomial functions which are generalisations of target concepts in the concept space are representable as hypotheses. More significantly for the learning behaviour, where it was noted above (p. 24) that in general for $k < k'$, $H_{M_{N,k}}^{CB1(\sigma_H)} \not\subseteq H_{M_{N,k'}}^{CB1(\sigma_H)}$ and $H_{M_{N,k}}^{CB1(\sigma_H)} \not\supseteq H_{M_{N,k'}}^{CB1(\sigma_H)}$, the hypothesis spaces of $CB2$ do include each other as k increases. It can be shown $H_{M_{N,k}}^{CB2} \subseteq H_{M_{N,k'}}^{CB2}$ for $k \leq k'$. In the case of $H^{CB1(\sigma_H)}$, $H_{M_{N,k}}^{CB1(\sigma_H)}$ cannot include $H_{M_{N,k'}}^{CB1(\sigma_H)}$ for a larger k' since there are functions whose representation depends on positive exemplars for a target $t \in M_{N,k}$ which cannot be included in a case-base extensible to a more specific target concept $t' \in M_{N,k'}$. However, in the case of $CB2$, the zero weightings introduced into the similarity measure mean that any positive exemplars in the case-base are equivalent for the purposes of classification (see p. 33) so that any 'problematic' positive exemplars that could not be included in the case-base for some more specific target concept can be replaced by appropriate positive instances of the new target concept without changing the represented function. This is the intuition behind the result stated in Proposition 5.20, with Proposition 5.19 as a lemma to that result.

Proposition 5.19 *For any two case bases CB, $CB' \subseteq (D_N \times \{0,1\})$, and any two weighted similarity measures $\sigma_{\overline{w}}$ and $\sigma_{\overline{w}'}$ defined in terms of vectors $\overline{w}, \overline{w}' \in \{0,1\}^N$, then the functions represented by $\langle CB, \sigma_{\overline{w}} \rangle$ and $\langle CB', \sigma_{\overline{w}'} \rangle$ are equivalent:*

$$h_{\langle CB,\sigma_{\overline{w}}\rangle} = h_{\langle CB',\sigma_{\overline{w}'}\rangle}$$

provided that **1)** *for any bit of the representation, $\overline{w}$ differs from $\overline{w}'$ only where $\overline{w}$ has value 0 on that bit while $\overline{w}'$ has value 1, and in addition all descriptions d' such that $(d', n) \in CB'$ also all agree on that bit (equation (111)),* **2)** *for all exemplars $(d, n) \in CB$ there is some corresponding $(d', n) \in CB'$ s.t. d and d' differ only where $(w)_i \neq (w')_i$ (and $(w)_i = 0$) (equation (112)),* **3)** and vice versa... *for all exemplars $(d', n) \in CB'$ there is some corresponding $(d, n) \in CB$ s.t. d and d' differ only where $(w)_i \neq (w')_i$ (equation (113)).*

$$\forall 1 \leq i \leq N \cdot (w)_i \neq (w')_i \rightarrow [(w)_i = 0 \wedge \exists b \in \{0,1\} \cdot \forall (d', n) \in CB' \cdot (d')_i = b] \tag{111}$$
$$\forall (d, n) \in CB \cdot \exists (d', n) \in CB' \cdot \forall 1 \leq i \leq N \cdot (w)_i = (w')_i \rightarrow (d)_i = (d')_i \tag{112}$$
$$\forall (d', n) \in CB' \cdot \exists (d, n) \in CB \cdot \forall 1 \leq i \leq N \cdot (w')_i = (w)_i \rightarrow (d')_i = (d)_i \tag{113}$$

Proof: Assume that the conditions stated above hold. We need to show $h_{\langle CB,\sigma_{\overline{w}}\rangle} = h_{\langle CB',\sigma_{\overline{w}'}\rangle}$ or $\forall d \in D_N \cdot h_{\langle CB,\sigma_{\overline{w}}\rangle}(d) = 1 \leftrightarrow h_{\langle CB',\sigma_{\overline{w}'}\rangle}(d) = 1$. Hence (equation (4)) proof will be established by showing:

$$\forall d \in D_N \cdot (\exists (d_{pos}, 1) \in CB \cdot \forall (d_{neg}, 0) \in CB \cdot \sigma_{\overline{w}}(d, d_{pos}) > \sigma_{\overline{w}}(d, d_{neg})) \leftrightarrow$$
$$(\exists (d'_{pos}, 1) \in CB' \cdot \forall (d'_{neg}, 0) \in CB' \cdot \sigma_{\overline{w}'}(d, d'_{pos}) > \sigma_{\overline{w}'}(d, d'_{neg}))$$

a) Only If Assume first that an instance d is classified positively by $h_{\langle CB,\sigma_{\overline{w}}\rangle}$ i.e. $\exists (d_{pos}, 1) \in CB \cdot \forall (d_{neg}, 0) \in CB \cdot \sigma_{\overline{w}}(d, d_{pos}) > \sigma_{\overline{w}}(d, d_{neg})$. Let d_1 be the descriptor of a positive exemplar fulfilling this proposition so that:

$$(d_1, 1) \in CB \tag{114}$$
$$\forall (d_{neg}, 0) \in CB \cdot \sigma_{\overline{w}}(d, d_1) > \sigma_{\overline{w}}(d, d_{neg}) \tag{115}$$

Let d'_1 be the descriptor of a positive exemplar in CB' satisfying equation (112) with respect to d_1. Hence:

$$(d'_1, 1) \in CB' \tag{116}$$
$$\forall 1 \leq i \leq N \cdot (w)_i = (w')_i \rightarrow (d_1)_i = (d'_1)_i \tag{117}$$

Consider a negative exemplar $(d'_2, 0) \in CB'$. Let d_2 be the descriptor of a negative exemplar in CB satisfying equation (113) with respect to d'_2. From equation (113) we have:

$$(d_2, 0) \in CB \tag{118}$$
$$\forall 1 \leq i \leq N \cdot (w')_i = (w)_i \rightarrow (d'_2)_i = (d_2)_i \tag{119}$$

From equation (111) we infer both that on those bits where $\overline{w}$ and $\overline{w}'$ disagree, we have $(w)_i = 0$ and $(w')_i = 1$ and also that any bits which are non-zero in $\overline{w}$ will also be non-zero in $\overline{w}'$. Additionally, d and d'_1 will agree with each other on these non-zero bits iff d and d_1 agree (equation (117)), ditto d, d_2 and d'_2 (equation (119)). Therefore if n is the number of bits s.t. $(w)_i = 1$, the number of bits on which $\overline{w}'$ is non-zero and on which d agrees with d'_1 is $(n \cdot \sigma_{\overline{w}}(d, d_1) + x_1)$, where x_1 is the number of bits s.t. $(w)_i = 0$, $(w')_i = 1$ and $(d)_i = (d'_1)_i$. Similarly the number of bits on which $\overline{w}'$ is non-zero and on which d agrees with d'_2 is $(n \cdot \sigma_{\overline{w}}(d, d_2) + x_2)$, where x_2 is defined by analogy to x_1. But since in addition, on any bits where $\overline{w}$ and $\overline{w}'$ differ then all exemplars in CB' will agree on that bit (equation (111)), $x_1 = x_2$. Therefore $\sigma_{\overline{w}}(d, d_1) > \sigma_{\overline{w}}(d, d_2) \leftrightarrow \sigma_{\overline{w}'}(d, d'_1) > \sigma_{\overline{w}'}(d, d'_2)$. Equation (115) requires $\sigma_{\overline{w}}(d, d_1) > \sigma_{\overline{w}}(d, d_2)$ and hence $\sigma_{\overline{w}'}(d, d'_1) > \sigma_{\overline{w}'}(d, d'_2)$. Since d'_2 is the description of any negative exemplar in CB', we have shown:

$$\forall (d'_2, 0) \,\epsilon\, CB' \cdot \sigma_{\overline{w}'}(d, d'_1) > \sigma_{\overline{w}'}(d, d'_2) \qquad (120)$$

Thus, any instance which is classified positively by the function represented by $\langle CB, \sigma_{\overline{w}} \rangle$ will also be classified positively by the function represented by $\langle CB', \sigma_{\overline{w}'} \rangle$. **b) If.** Assume instead that an instance d is classified positively by $h_{\langle CB', \sigma_{\overline{w}'} \rangle}$ i.e. $\exists (d'_{pos}, 1) \,\epsilon\, CB' \cdot \forall (d'_{neg}, 0) \,\epsilon\, CB' \cdot \sigma_{\overline{w}'}(d, d'_{pos}) > \sigma_{\overline{w}'}(d, d'_{neg})$. In analogy to above, there is therefore some positive exemplar in CB' with description d'_1 s.t. $\forall (d'_{neg}, 0) \,\epsilon\, CB' \cdot \sigma_{\overline{w}'}(d, d'_1) > \sigma_{\overline{w}'}(d, d'_{neg})$ and there is an exemplar $(d_1, 1) \,\epsilon\, CB$ which disagrees with d'_1 only where $\overline{w}$ has value 0 and $\overline{w}'$ has value 1. Consider a negative exemplar $(d_2, 0) \,\epsilon\, CB$; again (equation (112)), there is a negative exemplar $(d'_2, 0) \,\epsilon\, CB'$ whose description d'_2 only disagrees with d_2 where $\overline{w}$ and $\overline{w}'$ differ. As above, $\sigma_{\overline{w}}(d, d_1) > \sigma_{\overline{w}}(d, d_2) \leftrightarrow \sigma_{\overline{w}'}(d, d'_1) > \sigma_{\overline{w}'}(d, d'_2)$, since all exemplars $(d', n) \,\epsilon\, CB'$ agree on the bits which are not counted due to zero weightings in $\overline{w}$. Hence $\forall (d'_{neg}, 0) \,\epsilon\, CB' \cdot \sigma_{\overline{w}'}(d, d'_1) > \sigma_{\overline{w}'}(d, d'_{neg}) \rightarrow \forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_{\overline{w}}(d, d_1) > \sigma_{\overline{w}}(d, d_{neg})$, so we conclude that d is a positive instance of the function represented by $\langle CB, \sigma_{\overline{w}} \rangle$. □

Proposition 5.20 *The effective hypothesis space* $H^{CB2}_{M_{N,k}}$ *of* $CB2$ *with respect to the concept space* $M_{N,k}$ includes *the hypothesis space* $H^{CB2}_{M_{N,k'}}$ *with respect to* $M_{N,k'}$ *for any* $N \geq k' \geq k$.

$$\forall N \geq 1 \cdot \forall 1 \leq k \leq N \cdot \forall k \leq k' \leq N \cdot H^{CB2}_{M_{N,k}} \subseteq H^{CB2}_{M_{N,k'}}$$

Proof: Appealing to proposition 5.18 it is sufficient to show that for any values $1 \leq k \leq k' \leq N$:

$$\begin{aligned}&\forall f \,\epsilon\, B_N \cdot \\ &\quad (\exists t \,\epsilon\, M_{N,k} \cdot \exists CB \subseteq t \cdot h_{\langle CB, \sigma_{\overline{w}_t} \rangle} = f) \rightarrow (\exists t' \,\epsilon\, M_{N,k'} \cdot \exists CB' \subseteq t' \cdot h_{\langle CB', \sigma_{\overline{w}_{t'}} \rangle} = f)\end{aligned} \qquad (121)$$

Assume therefore there is some function $f \,\epsilon\, B_N$ which has a case-based representation $\langle CB, \sigma_{\overline{w}_t} \rangle$ s.t. $CB \subseteq t$ and t is a monomial function in $M_{N,k}$. Let U be the monomial representation of t and let U' be any monomial representation s.t. $U \subseteq U'$, $|U'| = k'$ and the set difference $U' \setminus U$ contains *only negative literals.*

Note also that, following our definition of monomial representations, we assume that neither U nor U' contains a pair of contradictory literals u_i and $\overline{u}_i$. Hence $u_i \,\epsilon\, U \rightarrow u_i \,\epsilon\, U'$, and $u_i \,\epsilon\, U' \rightarrow \overline{u}_i \not\epsilon U'$. Since also $\overline{u}_i \,\epsilon\, U \rightarrow \overline{u}_i \not\epsilon U' \setminus U$, we have:

$$\forall 1 \leq i \leq N \cdot \{u_i, \overline{u}_i\} \cap U \neq \emptyset \rightarrow \overline{u}_i \not\epsilon U' \setminus U \qquad (122)$$

Consider the following case-base CB' defined in terms of CB, U and U'.

$$\begin{aligned}CB' = \{&(d', n) \,\epsilon\, (D_N \times \{0, 1\}) | \\ &(\forall 1 \leq i \leq N \cdot \overline{u}_i \,\epsilon\, (U' \setminus U) \rightarrow (d')_i = 0) \wedge \\ &\exists (d, n) \,\epsilon\, CB \cdot (\forall 1 \leq i \leq N \cdot (d)_i \neq (d')_i \rightarrow \overline{u}_i \,\epsilon\, (U' \setminus U))\}\end{aligned} \qquad (123)$$

That is, CB' contains the exemplars $(d', n) \,\epsilon\, (D_N \times \{0, 1\})$ s.t. there is an exemplar (d, n) in CB whose description d agrees with d' on all bits unless the bit is constrained by a literal that appears in U' but not in U ($\overline{u}_i \,\epsilon\, U' \setminus U$) and, in addition, any such bit has the value 0 in the description d' ($(d')_i = 0$).

Clearly $t' = h_{U'}$ is a function in $M_{N,k'}$. In order to show that CB' satisfies equation (121), we must therefore show that $CB' \subseteq t'$ and that $h_{\langle CB', \sigma_{\overline{w}_{t'}}\rangle} = f = h_{\langle CB, \sigma_{\overline{w}_t}\rangle}$. **a)** $CB' \subseteq t'$**.** i.e. It must be shown that $(d', n) \epsilon CB' \rightarrow t'(d') = n$. For any $(d', n) \epsilon CB'$ equation (123) requires:

$$\forall (d', n) \epsilon CB' \cdot \forall 1 \leq i \leq N \cdot \overline{u}_i \epsilon (U' \setminus U) \rightarrow (d')_i = 0 \qquad (124)$$

$$\forall (d', n) \epsilon CB' \cdot \exists (d, n) \epsilon CB \cdot \forall 1 \leq i \leq N \cdot (d)_i \neq (d')_i \rightarrow \overline{u}_i \epsilon (U' \setminus U) \qquad (125)$$

Hence for a negative exemplar $(d'_{neg}, 0) \epsilon CB'$, let d_{neg} be the description of a corresponding negative exemplar in CB. Since $CB \subseteq t$, then clearly $t(d_{neg}) = 0$, and since U is the monomial representation for t, then:

$$\exists 1 \leq i \leq N \cdot (u_i \epsilon U \wedge (d_{neg})_i = 0) \vee (\overline{u}_i \epsilon U \wedge (d_{neg})_i = 1) \qquad (126)$$

Since $U' \supseteq U$ and hence any literal appearing in U will also appear in U', and since also for any value of i satisfying equation (126) then $\overline{u}_i \not\epsilon U' \setminus U$ (equation (122)) and hence by equation (125) d_{neg} and d'_{neg} will agree on that bit $((d_{neg})_i = (d'_{neg})_i)$. Thus there is a bit of d'_{neg} contradicting U', $\exists 1 \leq i \leq N \cdot (u_i \epsilon U' \wedge (d'_{neg})_i = 0) \vee (\overline{u}_i \epsilon U' \wedge (d'_{neg})_i = 1)$, and:

$$h_{U'}(d'_{neg}) = 0 \qquad (127)$$

Similarly, for a positive exemplar $(d'_{pos}, 1) \epsilon CB'$, there is a corresponding positive exemplar $(d_{pos}, 1) \epsilon CB$, and d_{pos} is a positive instance of h_U:

$$\forall 1 \leq i \leq N \cdot (u_i \epsilon U \rightarrow (d_{pos})_i = 1) \wedge (\overline{u}_i \epsilon U \rightarrow (d_{pos})_i = 0) \qquad (128)$$

As above, equations (122) & (125) require that for any bit constrained by a literal in U, then d_{pos} and d'_{pos} agree on those bits. In addition, for any additional negative literals appearing in $U' \setminus U$, equation (124) also requires that $(d'_{pos})_i = 0$, and so for any bit of the representation, $(u_i \epsilon U' \rightarrow (d'_{pos})_i = 1) \wedge (\overline{u}_i \epsilon U' \rightarrow (d'_{pos})_i = 0)$. Hence:

$$h_{U'}(d'_{pos}) = 1 \qquad (129)$$

b) $h_{\langle CB', \sigma_{\overline{w}_{t'}}\rangle} = h_{\langle CB, \sigma_{\overline{w}_t}\rangle}$**.** This can be shown by establishing that the preconditions of proposition 5.19 hold with respect to the case-bases CB and CB', and the weight vectors $\overline{w}_t$ and $\overline{w}_{t'}$. Assume $(w_t)_i \neq (w_{t'})_i$. Since $U \subseteq U'$, any literal in U also appears in U', and $\{u_i, \overline{u}_i\} \cap U \neq \emptyset \rightarrow (w_t)_i = (w_{t'})_i$ by equation (64). Hence neither literal can be in U, and $(w_t)_i = 0$. Additionally, equation (64) indicates that given $(w_t)_i \neq (w_{t'})_i$, either $u_i \epsilon U' \triangle U$ or $\overline{u}_i \epsilon U' \triangle U$. From the definition of U', the only possibility is $\overline{u}_i \epsilon U' \setminus U$, and equation (124) requires $(d')_i = 0$ for any $(d', n) \epsilon CB'$. Hence (111).

Assume $(d, n) \epsilon CB$. Define the descriptor d' s.t.

$$\begin{array}{rcl} (d')_i & = & 0 \;\; \textit{if} \;\; \overline{u}_i \epsilon U' \setminus U \\ (d')_i & = & (d)_i \;\; \textit{otherwise} \end{array}$$

Clearly, d' is a descriptor which disagrees with d only on bits i s.t. $\overline{u}_i \epsilon U' \setminus U$. Hence by equation (123), $(d', n) \epsilon CB'$. Assume for an index i that $(w_t)_i = (w_{t'})_i$; then by equation (64), $\{u_i, \overline{u}_i\} \cap U = \emptyset \leftrightarrow \{u_i, \overline{u}_i\} \cap U' = \emptyset$ and $\{u_i, \overline{u}_i\} \cap U' \setminus U = \emptyset$; thus, specifically, $\overline{u}_i \not\epsilon U' \setminus U$ and hence $(d')_i = (d)_i$. Hence (112).

Finally, take some $(d', n) \epsilon CB'$. Equation (125) requires the existence of some $(d, n) \epsilon CB$ s.t. $(d)_i \neq (d')_i$ only if $\overline{u}_i \epsilon U' \setminus U$. Assume then that $(w_t)_i = (w_{t'})_i$. As above, $\overline{u}_i \not\epsilon U' \setminus U$, giving $(d)_i = (d')_i$. Hence (113).

Thus by proposition 5.19, $h_{\langle CB', \sigma_{\overline{w}_{t'}}\rangle} = h_{\langle CB, \sigma_{\overline{w}_t}\rangle}$. □

Proposition 5.20 shows how the availability of knowledge about relevant and irrelevant features 'collapses' the hypothesis space of $CB2$ so that many of the irrelevant hypotheses output by $CB1(\sigma_H)$ are avoided entirely. For a fixed representation size, the size of the hypothesis space will increase monotonically with k as a direct corollary of Proposition 5.20. This gives an explanation of some kind for the contrast between the learning curves of Figures 4 & 5. What the result does not highlight, however, is the fact that, given the knowledge encoded in $\sigma_{\overline{w}_t}$, the learning behaviour of $CB2$ becomes independent of the representation size N. This is shown in the results of the following section.

Sample complexity of $CB2$ is independent of Representation Size

One claim that is easily established is that the sample complexity of $CB2$ increases with the number of relevant attributes k, but is independent of the size of the representation N. This is easily done by considering the 'covering net' technique applied to case-based learning algorithms by Albert and Aha [AA91].

Definition 5.9 ϵ-Net for CB2. *A set of instances $D \subseteq X$ is an ϵ-net for CB2 w.r.t a target concept t iff, except for a set of exceptions X' occurring with probability $< \epsilon$ ($X' \subseteq X$ s.t. $\mu X' < \epsilon$), for any instance $x \,\epsilon\, X$ there is an element of the net $d \,\epsilon\, D$ s.t. x and d agree on all bits relevant to the target concept. i.e. D is an ϵ-net iff:*

$$\mu\{x \,\epsilon\, X | \forall d \,\epsilon\, D \cdot \sigma_{\overline{w}_t}(d, x) < 1\} < \epsilon$$

The more elaborate method of constructing a bound on the sample complexity shown in [AA91] and [AKA91] is not necessary in the case of the *finite* example space D_N. In the current context, the definition of a covering net is sufficient in itself to guarantee a good hypothesis, so a bound on the sample complexity can be established without any further assumptions. Definition 5.9 is equivalent to requiring that each equivalence class in the partition $\langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle$ (p. 33) is 'hit' by some exemplar in the training sample, apart from exceptions occurring with sum total probability $< \epsilon$. Since all members of each of these classes are equivalent for the purposes of classification, then this is clearly sufficient to ensure correct classifications for the majority of the example space.

Proposition 5.21 *If the elements contained in a sample $\overline{x} \,\epsilon\, X^m$, $E_{\overline{x}} = \bigcup_{1 \leq i \leq m} \{x_i\}$, are an ϵ-net for $CB2$ w.r.t a target concept t, then the error in the hypothesis output by $CB2$ on $\overline{x}_t$ will be less than ϵ.*

$$\epsilon\text{-}net(E_{\overline{x}}) \to er_\mu(CB2(\overline{x}_t), t) < \epsilon$$

Proof: For some sample $\overline{x} \,\epsilon\, X^m$, assume that $E_{\overline{x}} = \bigcup_{1 \leq i \leq m} \{x_i\}$ is an ϵ-net for $CB2$ w.r.t t. Hence:

$$\mu\{x \,\epsilon\, X | \forall x' \,\epsilon\, E_{\overline{x}} \cdot \sigma_{\overline{w}_t}(x', x) < 1\} < \epsilon \tag{130}$$

Hence $\mu\{x \,\epsilon\, X | \exists x' \,\epsilon\, E_{\overline{x}} \cdot \sigma_{\overline{w}_t}(x', x) = 1\} > 1 - \epsilon$. Let the hypothesis output by $CB2$ on $\overline{x}_t$ be $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}$. Assume that for some example $x \,\epsilon\, X$, the element of the partition containing x is hit by a training example from $\overline{x}$: $\exists x' \,\epsilon\, E_{\overline{x}} \cdot \sigma_{\overline{w}_t}(x', x) = 1$ for some $x \,\epsilon\, X$. By the definition of $CB2$, $x' \,\epsilon\, E_{\overline{x}}$ implies $(x', t(x')) \,\epsilon\, CB$. Hence:

$$\exists (x', t(x')) \,\epsilon\, CB \cdot \sigma_{\overline{w}_t}(x', x) = 1 \tag{131}$$

Note that $\sigma_{\overline{w}_t}(x', x) = 1 \to t(x) = t(x')$, since x and x' agree on all relevant bits defining the target concept. Now assume $t(x) = 0$, so there must be a negative exemplar $(x', 0) \,\epsilon\, CB$ s.t. $\sigma_{\overline{w}_t}(x', x) = 1$, and hence immediately $\forall (d_{pos}, 1) \,\epsilon\, CB \cdot \exists (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_{\overline{w}_t}(x, d_{pos}) \leq \sigma_{\overline{w}_t}(x, d_{neg})$ and $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(x) = t(x) = 0$. Similarly, for any positive instance of t s.t. $t(x) = 1$, then we have a positive exemplar $(x', 1) \,\epsilon\, CB$ s.t. $\sigma_{\overline{w}_t}(x', x) = 1$. Additionally any negative exemplar will disagree with x on at least one relevant bit, giving $\forall (d_{neg}, 0) \,\epsilon\, CB \cdot \sigma_{\overline{w}_t}(x, x') = 1 > \sigma_{\overline{w}_t}(x, d_{neg})$. Hence $h_{\langle CB, \sigma_{\overline{w}_t} \rangle}(x) = t(x) = 1$. Thus $\forall x \,\epsilon\, X \cdot (\exists x' \,\epsilon\, E_{\overline{x}} \cdot \sigma_{\overline{w}_t}(x', x) = 1) \to CB2(\overline{x}_t)(x) = t(x)$. Hence $\mu\{x \,\epsilon\, X | \exists x' \,\epsilon\, E_{\overline{x}} \cdot \sigma_{\overline{w}_t}(x', x) = 1\} \leq \mu\{x \,\epsilon\, X | CB2(\overline{x}_t)(x) = t(x)\}$. Thus, $\mu\{x \,\epsilon\, X | CB2(\overline{x}_t)(x) = t(x)\} > 1 - \epsilon$ and $er_\mu(CB2(\overline{x}_t), t) = \mu\{x \,\epsilon\, X | CB2(\overline{x}_t)(x) \neq t(x)\} < \epsilon$. □

The probability of drawing an ϵ-net in a training sample of a fixed size is calculated exactly as in [AA91].

Proposition 5.22 c.f. [AA91, Lemma 2.1] [AKA91, Lemma 1] *The probability of drawing a sample which is not an ϵ-net for $CB2$ wrt $t \,\epsilon\, M_{N,k}$ is $< 2^k \cdot e^{-\frac{m\epsilon}{2^k}}$.*

$$\mu^m\{\overline{x} \,\epsilon\, X^m | E_{\overline{x}} \textit{is not an } \epsilon\textit{-net wrt } t \;\} < 2^k \cdot e^{-\frac{m\epsilon}{2^k}}$$

Proof: Let $G = \langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle$ be the partition induced on D_N by the similarity measure $\sigma_{\overline{w}_t}$ s.t. G contains 2^k regions of X and all descriptions within a $g \,\epsilon\, G$ agree on the k relevant bits of the description. Let G_1 be the subset of G s.t. $\forall g \,\epsilon\, G_1 \cdot \mu g \geq \frac{\epsilon}{2^k}$ and G_2 be the subset of G s.t. $\forall g \,\epsilon\, G_2 \cdot \mu g < \frac{\epsilon}{2^k}$.

Now, if for some $g \,\epsilon\, G$ there is an example in the sample which 'hits' g, then $\forall x \,\epsilon\, g \cdot \exists 1 \leq i \leq m \cdot \sigma_{\overline{w}_t}(x, x_i) = 1$. Hence so long as there is an example in the sample which hits every $g \,\epsilon\, G_1$, the probability of having an example $x \,\epsilon\, X$ s.t. $\forall 1 \leq i \leq m \cdot \sigma_{\overline{w}_t}(x, x_i) < 1$ is $\leq \sum_{g \,\epsilon\, G_2} \mu g < 2^k \cdot \frac{\epsilon}{2^k} = \epsilon$. Thus the probability of drawing a sample which is not an ϵ-net for $CB2$ wrt t is no more than the probability of drawing a sample $\overline{x} \,\epsilon\, X^m$ s.t. $\exists g \,\epsilon\, G_1 \cdot \forall 1 \leq i \leq m \cdot x_i \not\epsilon\, g$.

Taking one specific $g \,\epsilon\, G_1$, the probability that no example in the sample is taken from g is $\leq (1 - \frac{\epsilon}{2^k})^m$, and the probability that at least one $h \,\epsilon\, G_1$ is not hit by the sample can be no more than $|G_1| \cdot (1 - \frac{\epsilon}{2^k})^m$. Since $|G_1| \leq |G| = 2^k$, the result follows immediately from the inequality $(1 - x)^a < e^{-ax}$. $\square$

Hence:

Corollary 5.23 *The sample complexity of $CB2$ with respect to a concept space $M_{N,k}$ on k-literal monomials is no more than $\frac{2^k}{\epsilon} \log_e \frac{2^k}{\delta}$.*

$$m_{CB2}(M_{N,k}, \delta, \epsilon) \leq \frac{2^k}{\epsilon} \log_e \frac{2^k}{\delta}$$

Proof: From proposition 5.21, ϵ-net$(E_{\overline{x}}) \rightarrow er_\mu(CB2(\overline{x}_t), t) < \epsilon$, and from proposition 5.22 $\mu^m\{\overline{x} \,\epsilon\, X^m | E_{\overline{x}}$ is not an ϵ-net wrt $t \;\} < 2^k \cdot e^{-\frac{m\epsilon}{2^k}}$. Hence $\mu^m\{\overline{x} \,\epsilon\, X^m | er_\mu(CB2(\overline{x}_t), t) < \epsilon\} \leq \mu^m\{\overline{x} \,\epsilon\, X^m | E_{\overline{x}}$ is not an ϵ-net wrt $t \;\} < 2^k \cdot e^{-\frac{m\epsilon}{2^k}}$, and given large enough m, we can have $\mu^m\{\overline{x} \,\epsilon\, X^m | er_\mu(CB2(\overline{x}_t), t) < \epsilon\} < \delta$ for any value δ, given specifically by $m \geq \frac{2^k}{\epsilon} \log_e \frac{2^k}{\delta}$. $\square$

Hence it has been straightforward to establish an upper bound on the sample complexity of $CB2$ which is independent of the size of the representation, increasing only as a function of ϵ, δ and k. Corollary 5.23 is a poor bound, however, increasing as it does in $k.2^k$. For a tighter bound, the following section considers a better definition of a 'covering net', as used in the derivation of the sample complexity results of Blumer et al [BEHW89].

Upper bound on Sample Complexity of CB2

Definition 5.9, as used in the technique of Albert and Aha, requires the training sample to give an exhaustive coverage of the example space in order to guarantee a good approximation to the target concept. The neat, grid-like subdivision used by Albert and Aha has the disadvantage that it makes no allowance for the particulars of the learning algorithm or concept class being studied, and, therefore, many of the elements of the net are in fact redundant in learning the concept at hand. The general upper bounds on sample complexity in the PAC framework (i.e. those in terms of the VC dimension [BEHW89, Thm 2.1] and the 'Blumer bound' [BEHW87, Lemma 2.1] [BEHW89, Thm 2.2]), however, use a different notion of 'covering net' to bound the size of training sample needed, defined in terms of the effective hypothesis space of the learning algorithm being used. This depends on the notion of the 'error regions' of a learning algorithm. This is developed in the following statements, which differ from the formulation of [BEHW89] and [KV94] only in that the set of error regions $\Delta(t)$ is defined in terms of H_t, the set of hypotheses that might be output on training samples for a specific target concept, instead of the hypothesis space as a whole.

Definition 5.10 Error Regions *[KV94, p.57]. The error regions $\Delta(t)$ of a target concept t with respect to a hypothesis space H are the subsets of the example space on which a hypothesis in $H_t \subseteq H$, i.e. a hypothesis in H that might be output on a training sample for t, will disagree with the target concept t.*

$$\Delta(t) = \{h^{\Delta} \subseteq X | \exists h \,\epsilon\, H_t \cdot h^{\Delta} = \{x \,\epsilon\, X | h(x) \neq t(x)\}\}$$

Definition 5.11 ϵ-Error Regions *[KV94, p.57]. The ϵ-error regions $\Delta_{\epsilon}(t)$ of a target concept t with respect to a hypothesis space H are the error regions in $\Delta(t)$ for which the probability that a random example, drawn according to the fixed probability distribution, would fall into that region, is at least ϵ.*

$$\Delta_{\epsilon}(t) = \{r \,\epsilon\, \Delta(t) | \mu r \geq \epsilon\}$$

Definition 5.12 ϵ-Transversal *[BEHW89, p.952] [KV94, p.58]. An ϵ-transversal for a set of error regions $\Delta(t)$ is a set $S \subseteq X$ s.t. for every error region r with sum probability at least ϵ, there is some element $s \,\epsilon\, S$ which is also in r.*

$$\forall r \,\epsilon\, \Delta_{\epsilon}(t) \cdot S \cap r \neq \emptyset$$

Proposition 5.24 *[KV94, p.58] Consider an hypothesis space H, a target concept $t \,\epsilon\, B_N$ and a training sample $\overline{x}$ s.t. the set of exemplars contained in $\overline{x}$, $E_{\overline{x}} = \bigcup_{1 \leq i \leq m} \{x_i\}$, is an ϵ-transversal for the error regions of H w.r.t. t. For all consistent learning algorithms L using hypothesis space H, then the hypothesis output by L on $\overline{x}_t$ will have error $< \epsilon$.*

$$\textit{consistent}(L) \wedge \epsilon\textit{-transversal}(E_{\overline{x}}) \rightarrow er_{\mu}(L(\overline{x}_t), t) < \epsilon$$

Proof: Suppose that the training sample contains some example x_i s.t. $x_i \,\epsilon\, r$ for some ϵ-bad error region $r \,\epsilon\, \Delta_{\epsilon}(t)$. Then no hypothesis $h \,\epsilon\, H_t$ s.t. $h \triangle t = r$ can be output since this would be inconsistent with the example $(x_i, t(x_i))$. Hence, if the training sample contains an ϵ-transversal for $\Delta(t)$, then no ϵ-bad hypothesis can be output and $er_{\mu}(h, t) < \epsilon$. □

Proposition 5.24 gives a much better motivation for defining a covering net. Rather than cover some partition of the example space exhaustively, the ϵ-transversal requires exemplars only where they will specifically rule out some poor hypothesis. Hence definition 5.12 gives a sufficient condition over the training sample for a consistent learning algorithm to output an hypothesis with bounded error which is sensitive to the particular properties of the algorithm. To derive a bound on sample complexity, it is only necessary in addition to bound the probability of drawing a training sample containing an ϵ-transversal. Proposition 5.25 applies this in the case of $CB2$.

Proposition 5.25 *The probability of drawing a sample $\overline{x} \,\epsilon\, X^m$ whose examples do not form an ϵ-transversal for the error regions of $CB2$ with respect to a target concept $t \,\epsilon\, M_{N,k}$ is $< (|H_t| - 1)e^{-\epsilon m}$.*

$$\mu^m\{\overline{x} \,\epsilon\, X^m | E_{\overline{x}} \textit{ is not an } \epsilon\textit{-transveral for error regions of } CB2 \textit{ w. r. t. } t \,\epsilon\, M_{N,k}\} < (|H_t| - 1).e^{-\epsilon m}$$

Proof: As in [BEHW87, Lemma 2.1], take a specific $r \,\epsilon\, \Delta_{\epsilon}(t)$; since $\mu r \geq \epsilon$, the probability of a randomly drawn example not falling in r must be $< (1 - \epsilon)$, and the probability that no example in an m-sample falls in r is $< (1 - \epsilon)^m$. The event that some ϵ-bad error region contains no example in the training sample is the union of the events of each individual error region not being hit, so this probability is $< |\Delta_{\epsilon}(t)|.(1 - \epsilon)^m$. The result then follows since $|\Delta_{\epsilon}(t)| \leq |\Delta(t)|$, the error regions of $\Delta(t)$ are in one-to-one correspondence with the functions in H_t, and in general, $(1 - x)^a < e^{-ax}$. The '-1' appears since we exclude the function $t \,\epsilon\, H_t$ which has error 0. □

Corollary 5.26 *The sample complexity of $CB2$ with respect to the concept space $M_{N,k}$ is no more than $\frac{1}{\epsilon} \log_e \frac{(|H_t| - 1)}{\delta}$ for some $t \,\epsilon\, M_{N,k}$.*

$$m_{CB2}(M_{N,k}, \epsilon, \delta) \leq \frac{1}{\epsilon} \log_e \frac{(|H_t| - 1)}{\delta}$$

Proof: The event $er_\mu(CB2(\overline{x}_t, t) \geq \epsilon$ occurs only if $\overline{x}$ is not an ϵ-transversal for the error regions of $CB2$ with respect to a target concept $t \,\epsilon\, M_{N,k}$. Hence the probability of drawing a training sample resulting in an ϵ-bad hypothesis is bound by $(|H_t| - 1).e^{-\epsilon m}$ and $\mu^m\{\overline{x} \,\epsilon\, X^m | er_\mu(CB2(\overline{x}_t), t) \geq \epsilon\} < \delta$ provided that $\delta > (|H_t| - 1).e^{-\epsilon m}$. Hence result, solving for m as usual. □

Hence an upper bound on the sample complexity of $CB2$ can be derived directly from the characterisation of $CB2$-representable functions presented above (§5.1). This is demonstrated in the following subsection.

Counting $|H_t|$ for Upper Bound on Sample Complexity

Corollary 5.26 states an upper bound on the sample complexity of $CB2$ on the concept space $M_{N,k}$ in terms of $|H_t|$, the cardinality of the set of $CB2$-representable functions with respect to a target concept $t \,\epsilon\, M_{N,k}$ (equation (65)). Table 3 lists the functions that are $CB2$-representable with respect to target concepts defined in k-literals, $1 \leq k \leq 5$. The table was generated by systematically testing generalisations of representable formulae using the method of constructing and testing the canonical case-base $CB_{(f,t)}$ as indicated in Examples 5.6 and 5.7. Note that the property of '$CB2$-representability' will be preserved under transformations which *a)* permute the variables of a boolean form of the function and *b)* negate one or more of the variables. Those functions which are equivalent under permutation of the variables are said to be P-equivalent and those which are equivalent under permutation *and* negation are said to be NP-equivalent. For each NP-equivalence class of $CB2$-representable functions, Table 3 lists a representative function of the class, the size of the equivalence class and the size of the P-equivalence class for that representative function. The table is divided horizontally according to the value of k, the number of variables defining the functions. In counting the P- and NP-equivalence classes, only permutations and negations of the k defining variables are considered.

Table 4 collates the data from Table 3 to indicate the number of $CB2$-representable functions as a function of the size of the target concept k. The quantity $g(k)$ is the number of functions dependent on exactly k bits of the representation that are $CB2$-representable with respect to a fixed target concept $t \,\epsilon\, M_{N,k}$. This quantity is the sum of the sizes of the P-equivalence classes for a particular value of k listed in Table 3. Taking some function $t \,\epsilon\, M_{N,k}$, then H_t will only contain functions defined by the k bits defining the target concept t. For any such function defined in $k' \leq k$ variables, there are $\left(\begin{array}{c} k \\ k' \end{array}\right)$ ways of choosing k' from the k bits defining t. Hence:

$$|H_t| = \sum_{k'=0}^{k} \left(\begin{array}{c} k \\ k' \end{array}\right) g(k') \tag{132}$$

Table 4 also lists $|H^{CB2}_{M_{k,k}}|$, which is the number of functions that are $CB2$-representable with respect to *any* target concept defined on k out of k literals (all negations included). Clearly $|H^{CB2}_{M_{k,k}}|$ is a sum over the sizes of the NP-equivalence classes in Table 3, or alternatively:

$$|H^{CB2}_{M_{k,k}}| = \sum_{k'=0}^{k} 2^{k'} \left(\begin{array}{c} k \\ k' \end{array}\right) g(k') \tag{133}$$

The final column in Table 4 also lists for comparison the total number of functions definable on a k-dimensional example space, $|B_k| = 2^{2^k}$.

Hence combining the result of Corollary 5.26 with the values of $|H_t|$ listed in Table 4, the following upper bounds on sample complexity may be stated:

$$\forall N \geq 1 \cdot m_{CB2}(M_{N,1}, \delta, \epsilon) \leq \frac{1}{\epsilon} \log_e \frac{2}{\delta} < \frac{1}{\epsilon} + \frac{1}{\epsilon} \log_e \frac{1}{\delta} \tag{134}$$

k	Representative Function	No. P-equivalent functions	No. NP-equivalent functions
0	f_0^N	1	1
	f_1^N	1	1
1	u_1	1	2
2	u_1u_2	1	4
3	$u_1u_2u_3$	1	8
	$u_1u_2 + u_1u_3$	3	24
	$u_1u_2 + u_1u_3 + u_2u_3$	1	8
4	$u_1u_2u_3u_4$	1	16
	$u_1u_2 + u_1u_3u_4$	12	192
	$u_1u_2u_3 + u_1u_2u_4$	6	96
	$u_1u_2 + u_1u_3u_4 + u_2u_3u_4$	6	96
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_4$	4	64
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_4 + u_2u_3u_4$	1	16
5	$u_1u_2u_3u_4u_5$	1	32
	$u_1u_2 + u_1u_3u_4u_5$	20	640
	$u_1u_2u_3 + u_1u_4u_5$	15	480
	$u_1u_2u_3 + u_1u_2u_4u_5$	30	960
	$u_1u_2u_3u_4 + u_1u_2u_3u_5$	10	320
	$u_1u_2 + u_1u_3u_4u_5 + u_2u_3u_4u_5$	10	320
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5$	10	320
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_5$	60	1920
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_4u_5$	60	1920
	$u_1u_2u_3 + u_1u_4u_5 + u_2u_3u_4u_5$	15	480
	$u_1u_2u_3 + u_1u_2u_4u_5 + u_1u_3u_4u_5$	30	960
	$u_1u_2u_3u_4 + u_1u_2u_3u_5 + u_1u_2u_4u_5$	10	320
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4$	60	1920
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_5 + u_1u_4u_5$	15	480
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4u_5$	20	640
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_5 + u_2u_3u_4u_5$	60	1920
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_4u_5 + u_2u_3u_4u_5$	30	960
	$u_1u_2u_3 + u_1u_2u_4u_5 + u_1u_3u_4u_5 + u_2u_3u_4u_5$	10	320
	$u_1u_2u_3u_4 + u_1u_2u_3u_5 + u_1u_2u_4u_5 + u_1u_3u_4u_5$	5	160
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4 + u_1u_3u_5$	30	960
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4 + u_2u_3u_4$	30	960
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_3u_5 + u_1u_4u_5 + u_2u_3u_4u_5$	15	480
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4u_5 + u_2u_3u_4u_5$	10	320
	$u_1u_2u_3u_4 + u_1u_2u_3u_5 + u_1u_3u_4u_5 + u_1u_2u_4u_5 +$ $u_2u_3u_4u_5$	1	32
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4 + u_1u_3u_5 +$ $u_1u_4u_5$	5	160
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4 + u_1u_3u_5 +$ $u_2u_3u_4u_5$	30	960
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4 + u_1u_3u_5 +$ $u_2u_3u_4 + u_2u_3u_5$	10	320
	$u_1u_2u_3 + u_1u_2u_4 + u_1u_2u_5 + u_1u_3u_4 + u_1u_3u_5 +$ $u_1u_4u_5 + u_2u_3u_4 + u_2u_3u_5 + u_2u_4u_5 + u_3u_4u_5$	1	32

Table 3: $CB2$-representable functions for $1 \leq k \leq 5$

k	$g(k)$	$\lvert H_t \rvert \quad (t \,\epsilon\, M_{N,k})$	$\lvert H^{CB2}_{M_{k,k}} \rvert$	$\lvert B_k \rvert = 2^{2^k}$
0	2	2	2	2
1	1	3	4	4
2	1	5	10	16
3	5	13	60	256
4	30	62	674	65,536
5	603	820	22,148	4,294,967,296

Table 4: Numbers of $CB2$-representable functions

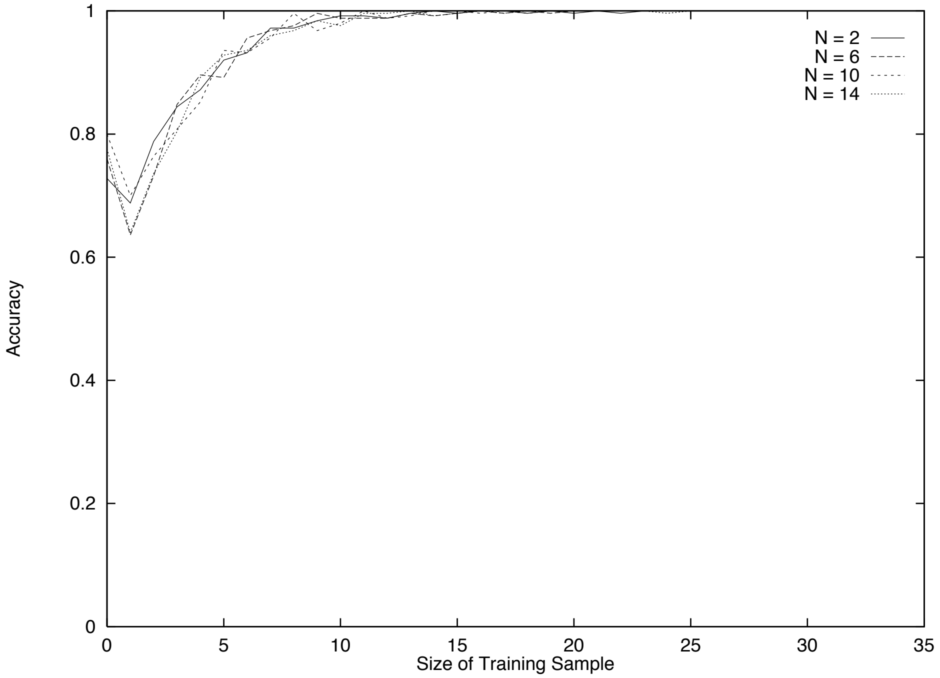

Figure 6: Average Learning Curve For Monomial Functions of 2 Literals by CB2

$$\forall N \geq 1 \cdot m_{CB2}(M_{N,2}, \delta, \epsilon) \leq \frac{1}{\epsilon} \log_e \frac{4}{\delta} < \frac{2}{\epsilon} + \frac{1}{\epsilon} \log_e \frac{1}{\delta} \tag{135}$$

$$\forall N \geq 1 \cdot m_{CB2}(M_{N,3}, \delta, \epsilon) \leq \frac{1}{\epsilon} \log_e \frac{12}{\delta} < \frac{3}{\epsilon} + \frac{1}{\epsilon} \log_e \frac{1}{\delta} \tag{136}$$

$$\forall N \geq 1 \cdot m_{CB2}(M_{N,4}, \delta, \epsilon) \leq \frac{1}{\epsilon} \log_e \frac{61}{\delta} < \frac{5}{\epsilon} + \frac{1}{\epsilon} \log_e \frac{1}{\delta} \tag{137}$$

$$\forall N \geq 1 \cdot m_{CB2}(M_{N,5}, \delta, \epsilon) \leq \frac{1}{\epsilon} \log_e \frac{819}{\delta} < \frac{7}{\epsilon} + \frac{1}{\epsilon} \log_e \frac{1}{\delta} \tag{138}$$

Conclusions: Learning Behaviour of CB2

In our definition of $CB2$ (Definition 5.2), we allowed the similarity measure to encode a privileged knowledge of which features are relevant to the definition of the chosen target concept. This has had two main areas of influence. On the one hand, reasoning about the case-based representation becomes a little more tractable; the represented function can clearly only depend on variables which are relevant

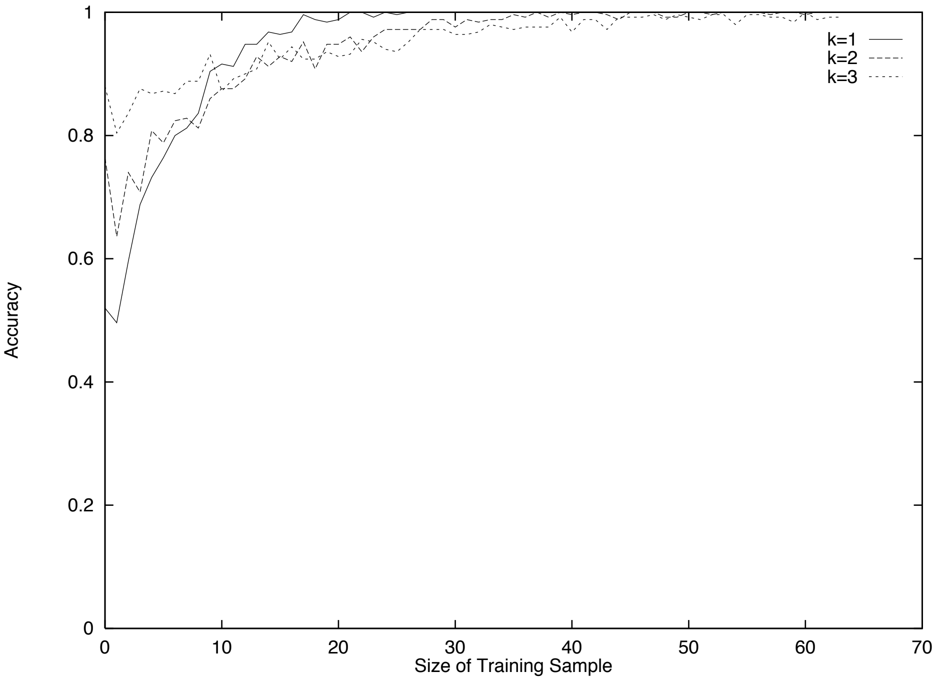

Figure 7: Average Learning Curve For Monomial Functions of k Literals by VS-CBR

to the target concept and will agree with the target concept on those variables. (See, for example, Proposition 5.15.) As a corollary of this, we have a simple method for testing whether a function is '$CB2$-representable' with respect to this perfectly weighted similarity measure $\sigma_{\overline{w}_t}$ (Proposition 5.5) and in addition we have the insight that representation with respect to $\sigma_{\overline{w}_t}$ is stable in that adding exemplars to a case-base representing the target concept will only ever produce a new representation of the target concept (p. 37).

On the other hand, the extra knowledge available to the similarity measure has a very marked impact on the learning behaviour of $CB2$ compared to the 'knowledge poor' $CB1(\sigma_H)$. Proposition 5.20 shows that for $k \leq k'$, $H^{CB2}_{M_{N,k}} \subseteq H^{CB2}_{M_{N,k'}}$ and hence $|H^{CB2}_{M_{N,k}}| \leq |H^{CB2}_{M_{N,k'}}|$; in general, larger training samples are required as k increases before $CB2$ will output a good approximation to a target in $M_{N,k}$ (in contrast to the observed behaviour of $CB1(\sigma_H)$). In addition, since a 'covering net' of examples (p. 54) for $CB2$ is defined in terms of the partition $\langle D_N \backslash \approx_{\sigma_{\overline{w}_t}} \rangle$ induced by the perfectly weighted similarity measure $\sigma_{\overline{w}_t}$ (p. 33) rather than in terms of individual examples in D_N, then the sample complexity depends only on k and is independent of the representation size N (Corollary 5.23). Figure 6 shows the average case learning curve of $CB2$, generated as in §5.3, for target concepts defined by monomial expressions of exactly 2 literals, defined on an example space of dimensionality 2, 6, 10 and 14 respectively. The very close fit between the learning curves in each case suggests that the learning behaviour is indeed independent of N in all aspects, as well as just the formal upper bound of the sample complexity.

6 Conclusions

In sections 4 and 5, some properties of the case-based learning algorithms $CB1(\sigma_H)$, VS-CBR and $CB2$ were presented. In this final section we will review the insights into these learning algorithms that our work has offered. Firstly, we will consider what our analysis of the idealised $CB2$ allows us to say about the 'practical' case-based learning algorithm VS-CBR. We will then discuss how the comparison of these learning algorithms illustrates the trade-off between sample complexity and computational complexity in inductive learning. Finally, we will consider an additional 'trick' involving the similarity measure which demonstrates the continuity between case-based learning and other forms of inductive learning.

The 'practical' algorithm VS-CBR (Definition 5.1) combines two processes; on the one hand, cases are collected into the case-base to construct a case-based representation as in $CB1(\sigma_H)$ and $CB2$, and on the other hand the examples are used immediately to infer the appropriate weighting for the similarity measure $\sigma_{\overline{w}}$. Although these two processes carry on in parallel, comparing the empirical behaviour of VS-CBR with that of $CB1(\sigma_H)$ and $CB2$ gives some further indications of the relative efficiencies of these two processes. Figure 7 shows the average learning curve for VS-CBR on monomial target concepts $t \in M_{6,k}$, $1 \leq k \leq 3$. (The graph was generated by exactly the same procedure as the other empirical results described here.) The figure shows the learning curves starting from the same initial accuracies defined by the proportion of positive and negative exemplars as observed in Figures 4 and 5. The learning curves however achieve a close to perfect classification accuracy much more quickly than the learning curve for $CB1(\sigma_H)$ shown in Figure 4. Indeed Figure 7 shows the learning curves 'crossing over' in much the same way as Figure 5, with VS-CBR starting on a lower initial accuracy for target concepts in $M_{6,1}$ but approaching perfect accuracy more rapidly than the other curves. Convergence is slower for VS-CBR than $CB2$ however (note especially the difference between the horizontal scales in Figures 5 and 7). In the average case, Figure 7 suggests that VS-CBR requires 12-13 training examples before $> 95\%$ classification accuracy is achieved for target concepts in $M_{6,1}$, while target concepts in $M_{6,2}$ and $M_{6,3}$ require ~ 20 and ~ 25 examples respectively.

The comparison of Figures 4, 5 and 7 suggests that the process of inferring a monomial representation from positive instances of the target concept (c.f. inferring the correct weighting for $\sigma_{\overline{w}}$ in VS-CBR) is much more efficient (with respect to sample size) than constructing a case-based representation $\langle CB, \sigma_H \rangle$ from exemplars of the target concept (VS-CBR *vs* $CB1(\sigma_H)$). This is completely in agreement with our assessment of $CB1(\sigma_H)$ in §4. On the other hand, convergence of the monomial learning algorithm is slower than the process of constructing a representation $\langle CB, \sigma_{\overline{w}_t} \rangle$ given *a priori* the correct weighting for $\sigma_{\overline{w}}$, for the range of parameters considered here (VS-CBR *vs* $CB2$). In addition to the extra examples needed, the learning behaviour of VS-CBR will be sensitive to the size of the representation (N) where $CB2$ is independent of this quantity; in the case of VS-CBR, an increasing number of examples will be needed to rule out irrelevant bits as the size of the representation is increased. Langley and Iba claim in passing that "we have shown analytically that the number of training instances required for [the standard monomial learning algorithm] to achieve a given level of accuracy [in the average case] increases only with the logarithm of the number of irrelevant attributes"[LI93, p.893], that is depends only on the quantity $N - k$ and not on the number of relevant attributes k. [1]

In addition, a further tradeoff is illustrated by the comparison of $CB1(\sigma_H)$ and VS-CBR. Note the low computational cost of executing $CB1(\sigma_H)$; VS-CBR however introduces an extra computational burden as it infers the correct weighting for $\sigma_{\overline{w}}$. We suggest that this is an example of a general trade-off between sample complexity and computational complexity that has been discovered in the PAC-learning framework [Hau90, p.1103]. Pitt and Valiant [PV88] present an example where one choice of representation makes the task of learning k-term DNF formulae computationally intractable, while a different representation (representing a larger class of functions) makes the learning task feasible at the cost of enlarging the hypothesis space and increasing the sample complexity. In the case of VS-CBR and $CB1(\sigma_H)$ we argue on the other hand that sample complexity is being reduced at the cost of extra computational complexity.

In defining $CB2$ as a variant of $CB1(\sigma_H)$, we were interested in exploring the implementation of inductive bias in case-based learning. The intuition behind notions of 'prior knowledge' in learning has clearly been borne out by our comparison of these two learning schemes. At the extreme end of the spectrum of 'knowledge-poor' to 'knowledge-rich' methods, consider the similarity measure σ_U^M defined

[1] In contrast note that Valiant gives the worst-case sample complexity for learning the class of all monomial functions as $2h(N + \log_2 h)$ by the standard algorithm, where N is the size of the representation, $\epsilon = \delta$ and $h = \frac{1}{\epsilon} = \frac{1}{\delta}$ [Val84a, p.444].

in terms of a monomial expression U as follows:

$$\sigma_U^M(d_1, d_2) = \begin{cases} 1 & if \quad h_U(d_1) = h_U(d_2) \\ 0 & otherwise \end{cases} \tag{139}$$

A variant of $CB2$ equipped with a measure σ_U^M will be a learning algorithm with the strongest possible bias for learning one specific target concept from M_N (indeed the measure reflects precise knowledge of this target concept). This 'algorithm' will output a perfect representation of the target concept $t = h_U$ provided that the training sample contains at least one positive and one negative example, and is trivially proved to be PAC with sample complexity $\frac{1}{\epsilon} \log \frac{2}{\delta}$. In addition to being defined in terms of the monomial expression itself, equation (139) incorporates a better bias for learning monomial concepts than the definitions of σ_H (Defn 3.1) and $\sigma_{\overline{w}}$ (Defn 3.2) since it ensures that hypotheses output in the case-based representation represent only monomial functions. The 'feature counting' approach of σ_H and $\sigma_{\overline{w}}$ is very much a disadvantage for learning monomial functions since, for small numbers of examples, hypotheses represented using these measures are very likely to contain disjunction.

Finally, consider a modification of VS-CBR (Definition 5.1) which infers a hypothesis $h_{\langle CB, \sigma_U^M \rangle}$ rather than $h_{\langle CB, \sigma_{\overline{w}} \rangle}$. In this case the algorithm will be exactly congruent with the standard learning algorithm for monomial functions of [Val84a]; the hypothesis output by either algorithm will be identical for any training sample for a monomial target concept $t \,\epsilon\, M_N$. The only difference that remains between the two approaches is that the standard algorithm produces hypotheses in a representation that is much more efficiently evaluated than the implicit, case-based representation used by the variant of VS-CBR. It is clear however that there is continuity between case-based learning and other forms of inductive learning, and especially that most kinds of learning can be introduced into a case memory system of the kind we have modelled through appropriate manipulation of the similarity measure. We believe that the work reported here has contributed to showing that the questions and problems of learning in case-based reasoning systems are essentially the same as those in systems more frequently studied in the field of machine learning, and that these questions can be usefully dealt with by existing techniques of the field.

References

[AA91] M K Albert and D W Aha. Analyses of instance-based learning algorithms. In *AAAI-91: Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 553–558, 1991.

[AB92] M Anthony and N Biggs. *Computational Learning Theory.* Cambridge University Press, 1992.

[AKA91] D W Aha, D Kibler, and M K Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[BEHW87] A Blumer, A Ehrenfeucht, D Haussler, and M K Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

[BEHW89] A Blumer, A Ehrenfeucht, D Haussler, and M K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, October 1989.

[Bib95] Y Biberman. The role of prototypicality in exemplar-based learning. In Nada Lavrac and Stefan Wrobel, editors, *Machine Learning: ECML-95 (Proc. 8th European Conf. on Machine Learning, 1995)*, Lecture Notes in Artificial Intelligence 914, pages 77–91, Berlin, Heidelberg, New York, 1995. Springer Verlag.

[Cam92] R M Cameron-Jones. Minimum description length instance-based learning. In *Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence*, pages 368–373. World Scientific, 1992.

[CCK93] R T Chi, M D Chen, and M Y Kiang. Generalised case-based reasoning system for portfolio management. *Expert Systems with Applications*, 6(1):67–76, 1993.

[CS93] S Cost and S Salzberg. A weighted nearest neighbour algorithm for learning with symbolic features. *Machine Learning*, 10(1):37–66, March 1993.

[Dea95] A M Dearden. *The use of formal models in the design of interactive case memory systems.* PhD thesis, University of York, UK, 1995.

[DF86] W H E Day and D P Faith. A model in partial orders for comparing objects by dualistic measures. *Mathemetical Biosciences*, 78(2):179–192, 1986.

[Glo95] C Globig. Fallbasiertes repräsentieren und lernen von begriffsmengen ("Case-based representation and concept learning"). In *Fallbasiertes Schliessen - Grundlagen & Anwendungen, Workshop auf der 3. Deutschen Expertensystemtagung (XPS-95), University of Kaiserslautern, March 1995.* Report LSA-95-02, Centre for Learning Systems and Applications, University of Kaiserslautern, 1995.

[GW94] C Globig and S Wess. Symbolic learning and nearest-neighbour classification. In H-H Bock, W Lenski, and M M Richter, editors, *Information Systems and Data Analysis: Prospects, Foundations, Applications. Proceedings of the 17th Annual Conference of the Gesellschaft fur Klassification e.V. University of Kaiserslautern, March 3-5, 1993.* Springer-Verlag, 1994.

[Hau88] D Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 36:177–221, 1988.

[Hau90] D Haussler. Probably approximately correct learning. In *AAAI-90 Proceedings of the Eight National Conference on Artificial Intelligence, Boston, MA*, pages 1101–1108. American Association for Artificial Intelligence, 1990.

[Jan92] K P Jantke. Case-based learning and inductive inference. GOSLER report 08/92, FB Mathematik & Informatik, TH Leipzig, 1992.

[JL93] K P Jantke and S Lange. Case-based representation and learning of pattern languages. In *EWCBR-93 Working Notes of the first European Workshop on Case-Based Reasoning*, volume 1, pages 139–144. University of Kaiserslautern, 1993.

[KV94] M J Kearns and U V Varizani. *An Introduction to Computational Learning Theory.* MIT Press, 1994.

[LI93] P Langley and W Iba. Average-case analysis of a nearest neighbour algorithm. In R Bajcsy, editor, *IJCAI-93: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 889–894. Morgan Kaufmann, 1993.

[Mur71] S Muroga. *Threshold Logic and its Applications.* John Wiley & Sons, 1971.

[Nat91] B K Natarajan. *Machine Learning: A Theoretical Approach.* Morgan Kaufmann, 1991.

[PV88] L Pitt and L G Valiant. Computational limitations on learning from examples. *Journal of the ACM*, 35(4):965–984, October 1988.

[Sch94] C Schaffer. A conservation law for generalization performance. In *ML94: Proceedings of the International Conference on Machine Learning, New Brunswick, New Jersey*, pages 259–265. Morgan Kaufmann, 1994.

[SDHK95] S Salzberg, A Delcher, D Heath, and S Kasif. Best-case results for nearest neighbour learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(6):599–608, June 1995.

[Tur93] P Turney. Theoretical analyses of cross-validation error and voting in instance-based learning. Technical Report NRC-35073, Knowledge Systems Laboratory, Institute for Information Technology, National Research Council (Canada), 1993.

[Utg86] P E Utgoff. Shift of bias for inductive concept learning. In R S Michalski, J G Carbonell, and T M Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach (Volume 2)*. Morgan Kaufmann, 1986.

[Val84a] L G Valiant. Deductive learning. *Philosophical Transactions of the Royal Philosophical Society of London A*, 312:441–446, 1984.

[Val84b] L G Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

[WG94] S Wess and C Globig. Case-based and symbolic classification algorithms - A case study using version space. In S Wess, K-D Althoff, and M M Richter, editors, *Topics in CBR: Selected papers from the First European Workshop on Case-Based Reasoning - EWCBR-93, Kaiserslautern, Germany, November '93*, Lecture Notes in Computer Science vol. 837, pages 77–91. Springer-Verlag, 1994.

[YJL94] Yasubumi S, K P Jantke, and S Lange. Learning languages by collecting cases and tuning parameters. In *ALT94: Proceedings of Fifth International Conference on Algorithmic Learning Theory*, 1994.

[Zha92] J Zhang. Selecting typical instances in instance-based learning. In *ML92: Proceedings of the 9th International Conference on Machine Learning*, pages 470–479. Morgan Kaufmann, 1992.